

# STA-380 Exercises

Khyati Jariwala, Aishwarya Rajeev, Juhi Patel, Varun Kausika

2022-08-10

Link to the GitHub: <https://github.com/VarunKausika/STA-380-Exercises>

## 1. Probability practice

### Part A

**Given information:**

Two categories of users:

1. Truthful clicker (TC)
2. Random clicker (RC)

**Information on probabilities:**

- $P(RC) = 0.3$
- $P(Yes|RC) = 0.5$
- $P(No|RC) = 0.5$
- $P(TC) = 0.7$
- $P(Yes|TC) = x$
- $P(No|TC) = 1 - x$
- $P(Yes) = 0.65$
- $P(No) = 0.35$

Using the Rule of Total Probability,

$$P(Yes) = P(Yes, TC) + P(Yes, RC) = P(TC) * P(Yes|TC) + P(RC) * P(Yes|RC) \quad (1)$$

$$P(Yes) = 0.7x + 0.3 * 0.5 = 0.7x + 0.15 = 0.65$$

Solving for x, we get,

$$x = P(Yes|TC) = 0.714$$

### Part B

We are being asked  $P(Diseased|Positive)$

**Given information:**

- $P(Positive|Diseased) = 0.993$

- $P(Negative|NotDiseased) = 0.9999$
- $P(Diseased) = 0.000025$

According to Bayes Rule and Rule of Total Probability,

$$P(Diseased|Positive) = \frac{P(Positive|Diseased) * P(Diseased)}{P(Positive)} \quad (2)$$

and,

$$P(Positive) = P(Positive|Diseased) * P(Diseased) + P(Positive|NotDiseased) * P(NotDiseased) \quad (3)$$

Therefore,

$$P(Positive) = 0.993 * 0.000025 + 0.0001 * 0.999975 = 0.000125$$

Substituting in (2) we get,

$$P(Diseased|Positive) = \frac{0.993 * 0.000025}{0.000125} = 0.1986$$

## 2. Wrangling the Billboard Top 100

### Part A

First, we load in the data and perform a group by on the performer and the song, with an agg function of count for the week

Table 1: Billboards

performer	song	count
'N Sync	(God Must Have Spent) A Little More Time On You	22
'N Sync	Bye Bye Bye	23
'N Sync	Gone	24
'N Sync	I Drive Myself Crazy	12
'N Sync	I Want You Back	24
'N Sync	It's Gonna Be Me	25
'N Sync	Pop	15
'N Sync	Tearin' Up My Heart	1
'N Sync	This I Promise You	26
'N Sync & Gloria Estefan	Music Of My Heart	20

Finally, we sort the dataframe in descending order of counts and find the top 10 and give our table a caption:

Table 2: Top 10 most popular songs

performer	song	count
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79

performer	song	count
Jason Mraz	I'm Yours	76
The Weeknd	Blinding Lights	76
LeAnn Rimes	How Do I Live	69
LMFAO Featuring Lauren Bennett & GoonRock	Party Rock Anthem	68
OneRepublic	Counting Stars	68
Adele	Rolling In The Deep	65
Jewel	Foolish Games/You Were Meant For Me	65
Carrie Underwood	Before He Cheats	64

## Part B

First we group by year and order by ascending year. Then, we remove the years 1958 and 2021 from the rows and order just to make sure. Finally, we proceed to plot the columns.

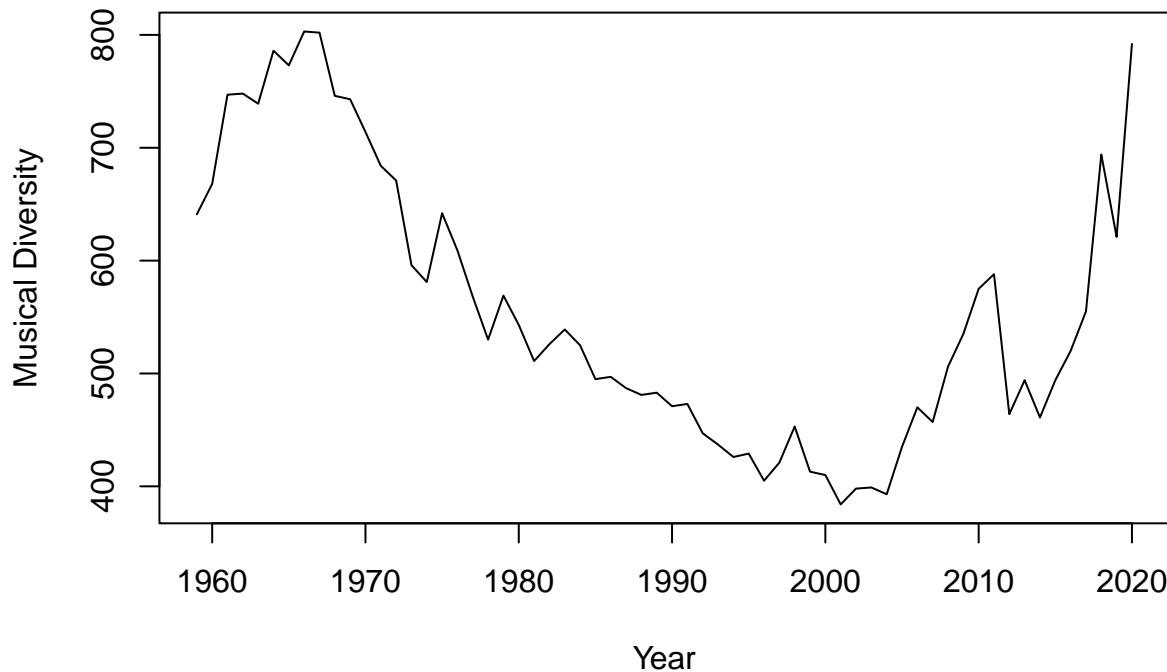


Figure 1: The plot shows peaks in diversity in 1965 and 2020, along with extreme lows in 2000

## Part C

First, we filter the dataframe from part A to include only those songs with weeks at least 10. Then, We do a group by on the artists. Finally, we can select those artists with a hit-count of at least 30. ??

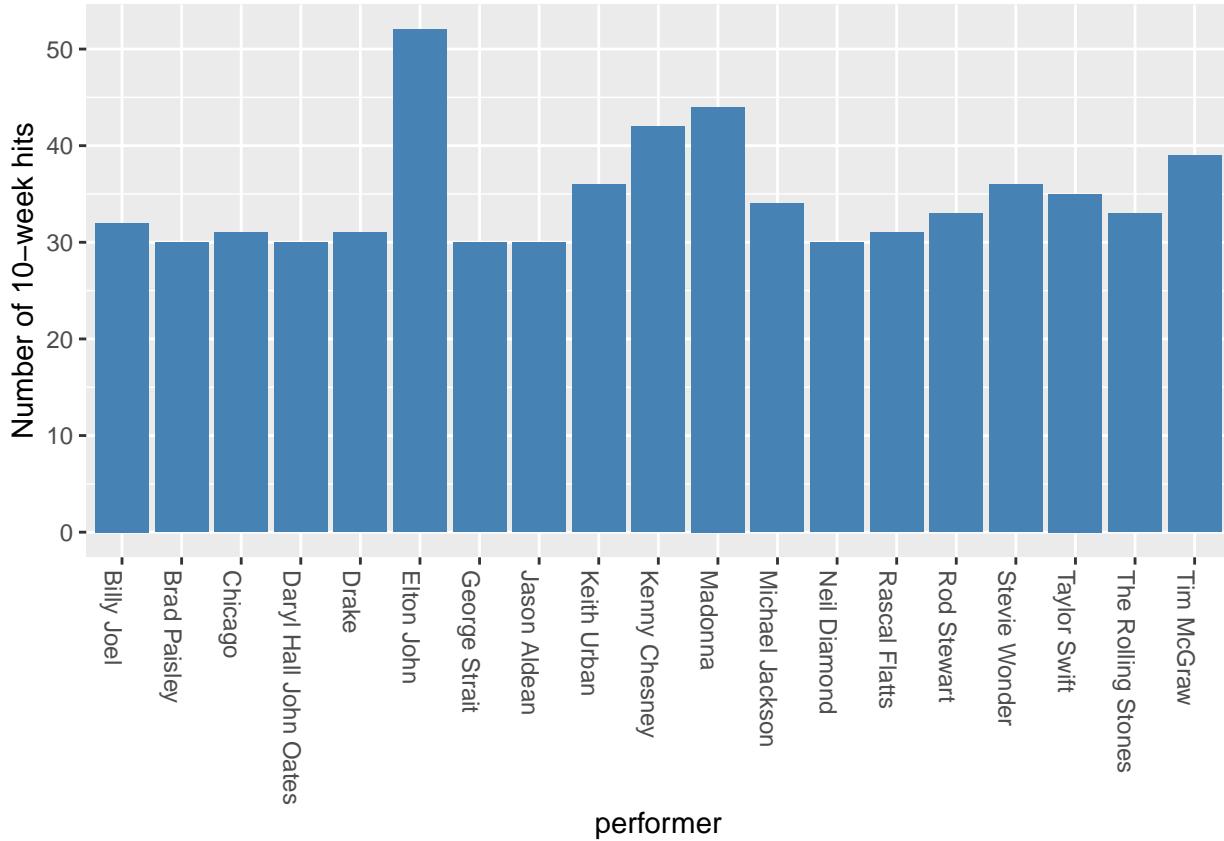


Figure 2: Elton John has more hits than others by quite a large margin

### 3. Visual Storytelling Part 1: green buildings

#### Outliers:

We are interested in finding the potential economic gains the owner could make by constructing a 15-story green building in the neighbourhood of East Cesar Chavez.

First, we remove the outliers from the dataset as mentioned by the analyst.

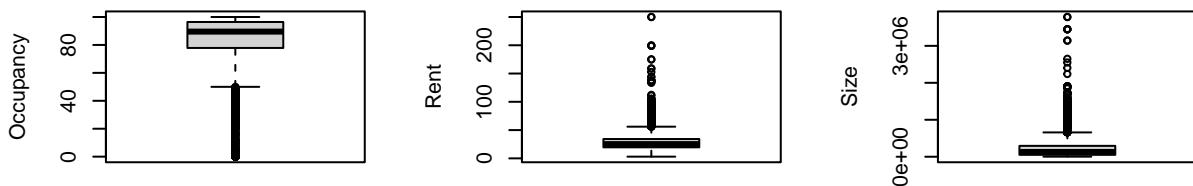


Figure 3: It seems like occupancy below 40 percent can be removed in contrast to the 10 percent that was suggested

### Observations:

1. Removing outliers in occupancy below 40% would only remove 456 rows from our dataset, so we proceed with this step.
2. We should not remove outliers from rent, as that would shrink our dataset greatly, and also since most of the green buildings have higher rents, this would remove them entirely.
3. We should not remove outliers from size, because size and rent are positively correlated and this might remove a significant number of green buildings as well.

### Correlation Matrix:

Now, we plot the feature correlation matrix in our dataset.

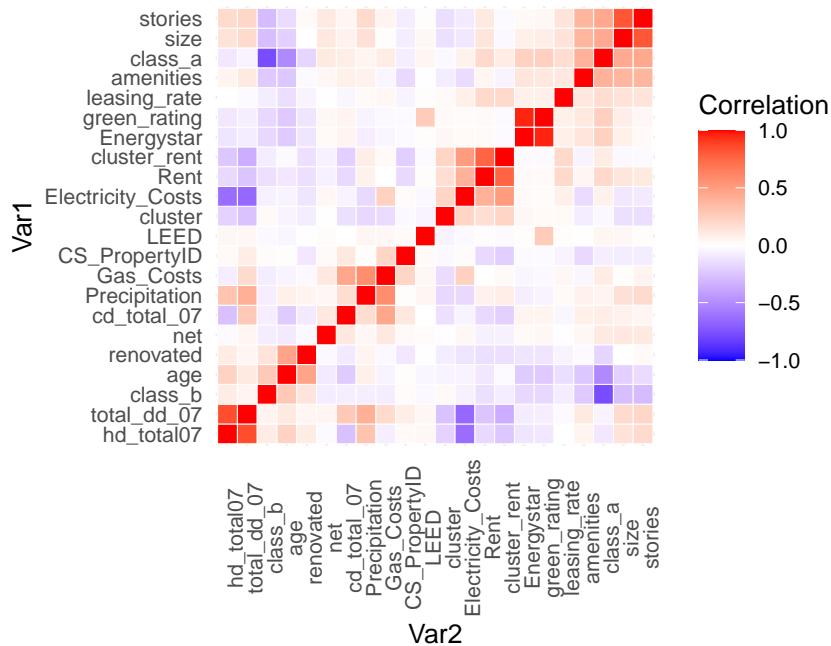


Figure 4: As we can see, rent is highly dependent on clusters

### Observations:

1. Rent and Cluster has a high positive correlation, because some neighbourhoods are more expensive than others. So, instead of taking the median rent throughout our dataset, we can just use the cluster rent variable.
2. Rent and age has a mildly negative correlation, as one would expect.
3. Rent is not correlated with green rating, LEED or Energy Star.
4. Class A buildings are positively correlated with rent whereas Class B buildings are negatively correlated.
5. Leasing rate is positively correlated with rent.
6. Size and stories(can be considered correlated features) are both positively correlated to rent.

### Scatter plots:

Now, we draw some scatter plots to understand the relation between correlated variables.

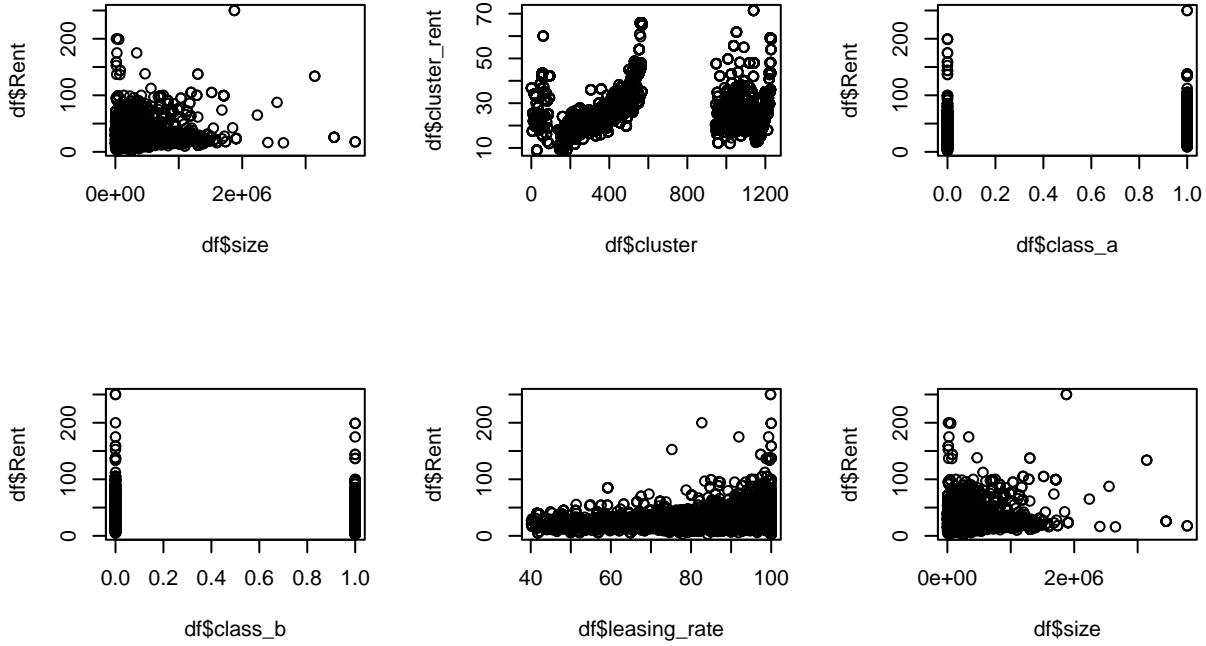


Figure 5: We can see clear patterns in the cluster rent v cluster. Some clusters have a significantly higher average rent.

## Observations

1. Rent in Class A buildings occur more in positive class. This could be because of the positive correlation of class A buildings with green rating, size and stories.
2. Rent in Class B buildings occur more in negative class. The correlation of Class B with the variables is exactly opposite to that of Class A.
3. This implies that the builder **MUST** aim to construct a Class A building in order to be able to charge more rent.
4. Furthermore, the builder might want to move the building location to a cluster that falls in higher rent bracket.

## Density plots:

Now, we look at the green buildings in our dataset and compare them to all the other buildings

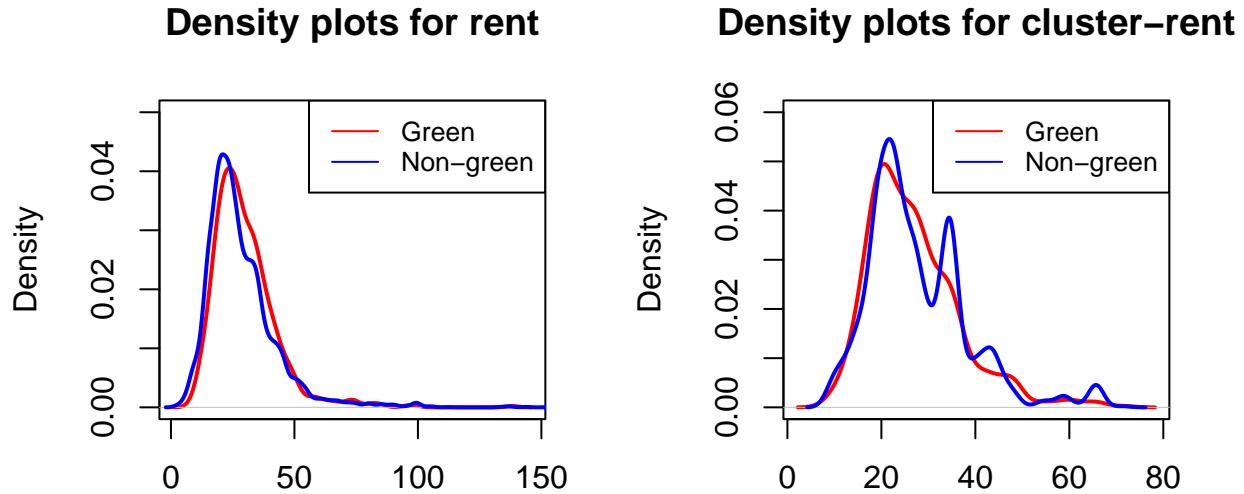


Figure 6: We can see that both green buildings and non-green buildings are not normally distributed. They are shifted to the left.

#### Observations:

1. We can conclude that this data is skewed toward the lower rent buildings (has long tails).
2. When comparing green and non-green buildings, the green buildings seem to peak at a slightly higher rent than the non green rents. Comparing the median values of both of these will be appropriate for the graph on the left.
3. However, for the cluster rent graph, we see that the peak of the non-green occurs at a slightly higher rent value than the peak for the green buildings. Therefore, if we use cluster rent as our primary variable to determine income (as we said we would before), we would expect in fact a lower rent value for the greens.
4. The cluster rent density function is not normal for either the greens or the non greens. For the non-greens, it looks more like a sum of many different bell shaped curves. This suggests that each cluster within the data has a separate normal distribution.

#### Economic Impact of Green Houses:

Economic impact could include a variety of factors in this case, namely:

1. **Sources of income depend on the rent being charged. Which inturn are caused by:**
  - (a) Size of the house being rented (250,000 sqft in our case).
  - (b) The neighbourhood of the house. A groupby statement is in order to find out the prices in East Cesar Chavez. On top of this, the premium rent for being green, as shown in the density plot, is pretty much no-existent. So controlling for the clusters, we can say that green houses don't do much better than non-green houses in this dataset.
  - (c) Age of the building. One would expect that the older the building is the lower the rent is. This is shown in the correlation plot.

- (d) Appreciation of value of houses in the neighbourhood. This is something that cannot be measured by the given features. A thorough analysis of previous time series data of prices is recommended.
- (e) Occupancy of the building (relevant variable here is lease.rate).
- (f) Whether amenities are available or not.
- (g) Whether it was renovated or not.
- (h) Whether it is Class A or Class B.

## 2. Sources of expenditure depend on:

- (a) Initial capex (100,000 in our case).
- (b) Premium needed to be spent on constructing a green building (5%).
- (c) Maintenance, repairs and other charges.
- (d) **NOTE: Water, Gas and Electricity charges are assumed to be paid by the tenant, hence are not included.**

### Where the Analyst went wrong:

The analyst found the impact on income in a very linear way. He did not take into account the following:

1. That the green buildings do not have any premium in rent after accounting for the cluster in which the buildings are.
2. That 40% of the occupancies are outliers, since they have a very high mean and low interquartile range.
3. Time value of money: he has calculated the simple Payback Period of the project. The more accurate estimator of success would have been the Net Present Value.
4. A cross sectional analysis across other buildings should have been done to compare Payback Periods instead of assuming 8 years is a decent time to recuperate costs
5. The worst case of occupancy is not 90%. It is much lower. As seen in the dataset many buildings have occupancies even lower than 40%.
6. That the cost of the buildings in the area may reduce/increase over time.
7. Accounting for other variables which may inhibit the ability to charge high rent (eg. Class A vs B, others listed above).

Confounding variables:

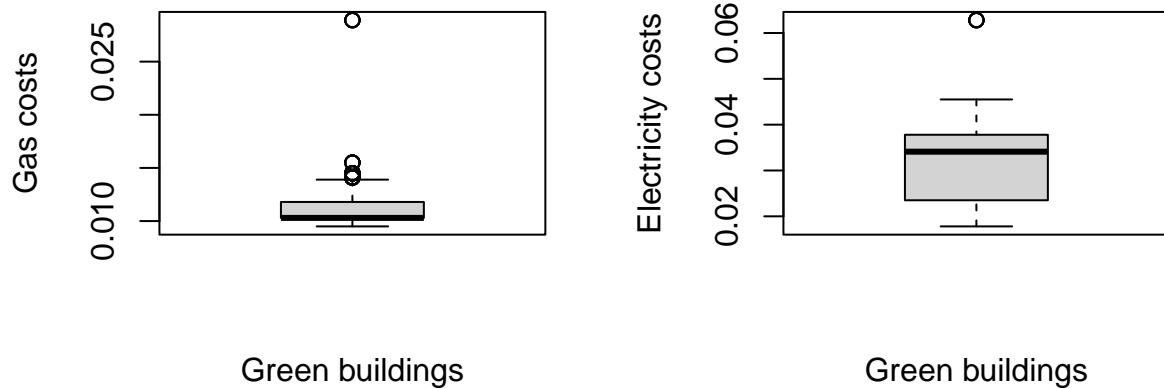


Figure 7: Some green houses show a high electricity and gas cost which goes against our intuition of what is green. The outliers can be adjusted for by removing them

#### 4. Visual Storytelling part 2: Capital Metro data:

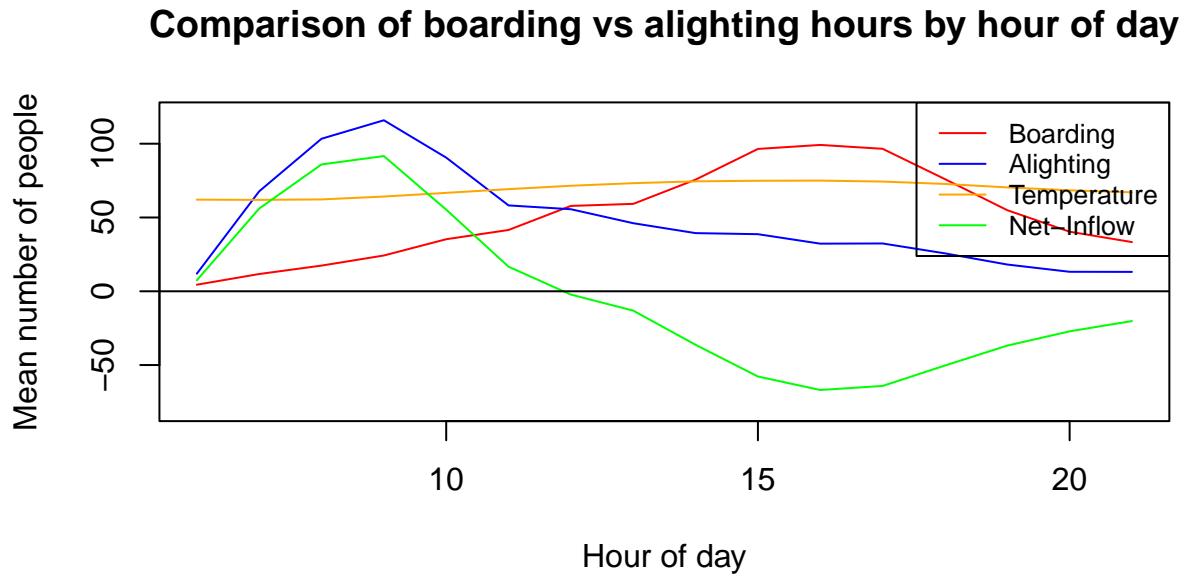


Figure 8: In the campus vicinity, alighting is most common in the morning (when classes start) and boarding is most common in the evening (when classes end). Furthermore, temperature doesn't seem to have an affect on the boardings and alightings. Here, net-inflow is calculated as alightings minus boardings. Net inflow is 0 around noon (break between classes).

### Comparison of boarding vs alighting means by month and hour of day

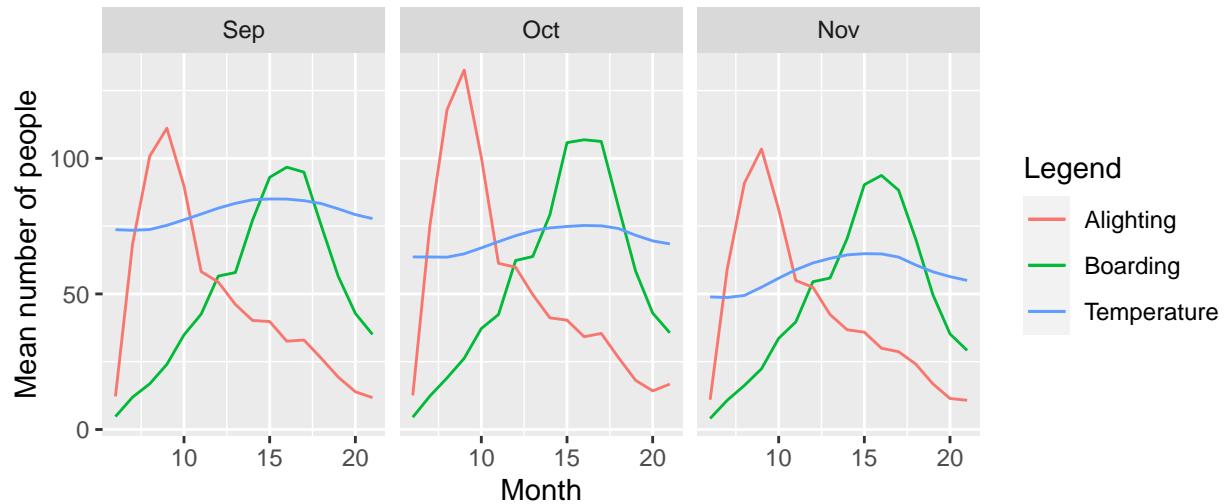


Figure 9: Similar patterns are seen across months, temperatures are highest in September and gradually come down

### Comparison of boarding vs alighting means by day of week

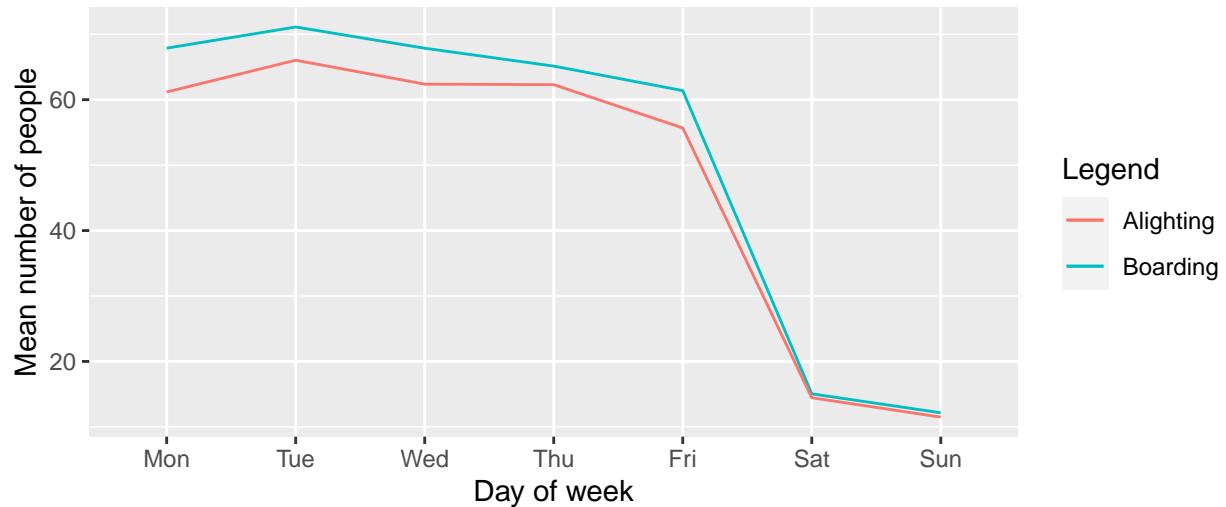


Figure 10: As we can see, the number of people boarding and alighting near campus on weekends are much lower.

## Comparison of boarding vs alighting means by month and day of week

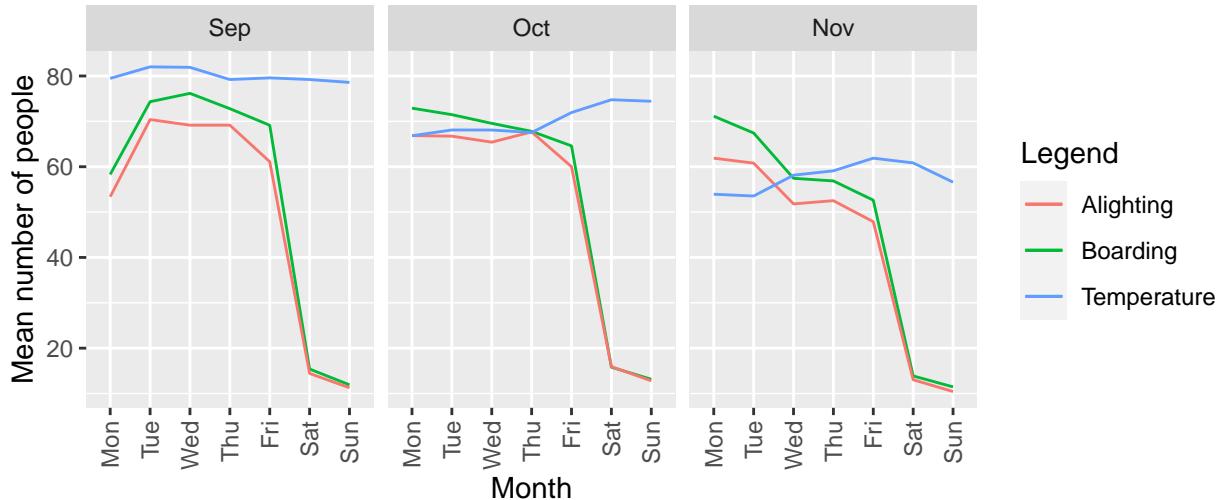


Figure 11: Similar patterns are seen across months, however significantly less people seem to attend classes on Monday in September.

## Comparison of boarding vs alighting means by month

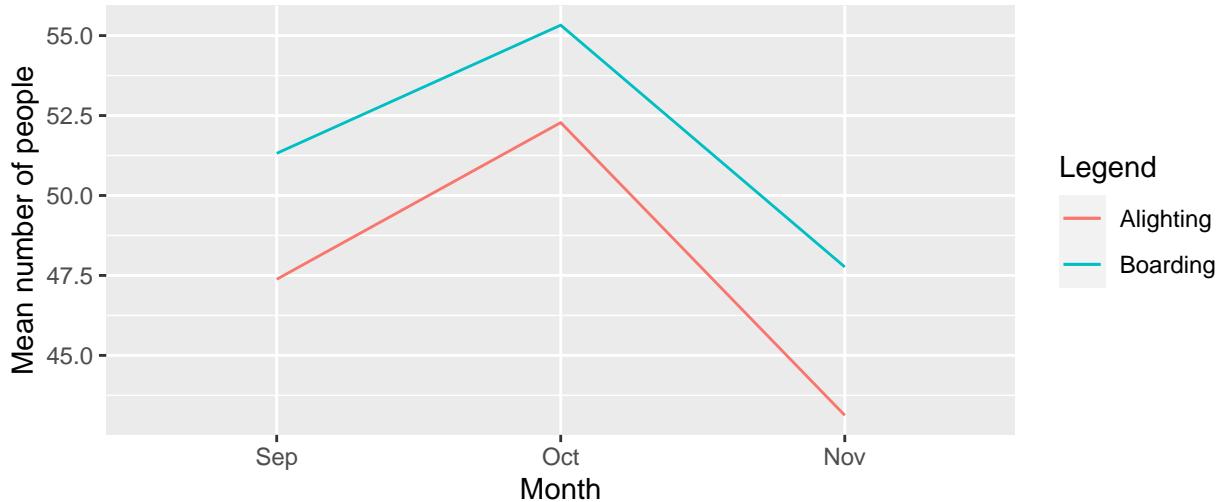


Figure 12: There seem to be more students coming in October compared to November and September. This is because of Thanksgiving holidays in November and since college starts in August, students might not immediately be using the buses on a regular basis in September.

## 5. Portfolio modeling

### Introduction

We chose the below ETFs to provide portfolio diversification and a range of risk.

- Invesco (QQQ): One of the biggest, exclusively non-financial stock, and heavily tech-heavy trusts.

- SPY: One of the safest and largest ETF.
- iShares Russell 1000 Growth ETF (IWF): One of the most popular US large-cap growth ETFs with a long track record. It is an aggressive growing ETF.
- SCO: A low performing stock.

In total, we have selected 3 ETFs - “QQQ”, “SPY”, “IWF”. We looked at data from ETFs for five years commencing on January 1, 2017

## Sample Data for QQQ

Table 3: QQQ Sample Data

QQQ.Open	QQQ.High	QQQ.Low	QQQ.Close	QQQ.Volume	QQQ.Adjusted
114.6697	115.3619	114.3043	114.9293	22307600	114.9293
115.0542	115.7657	115.0446	115.5542	19749100	115.5542
115.4677	116.2849	115.4677	116.2080	20644300	116.2080
116.3330	117.5347	116.0349	117.2271	24074300	117.2271
117.3232	117.8232	117.2463	117.6117	18909200	117.6117
117.6597	118.2270	117.3905	117.8712	16176600	117.8713
117.9193	118.1885	117.3136	118.1885	20686000	118.1885
117.7655	118.0828	116.9098	118.0058	19622500	118.0059
118.0635	118.6115	118.0635	118.4096	16930800	118.4096
118.1020	118.2847	117.7078	118.0539	14538500	118.0539

## Combine all the returns in a matrix

Table 4: All Returns

	C1C1.QQQa	C1C1.SPYa	C1C1.IWFa	C1C1.SCOa
2017-01-04	0.0054375	0.0059492	0.0077585	-0.0220811
2017-01-05	0.0056577	-0.0007945	0.0015961	-0.0204144
2017-01-06	0.0087697	0.0035778	0.0064680	0.0026839
2017-01-09	0.0032806	-0.0033009	-0.0000932	0.0606204
2017-01-10	0.0022071	0.0000000	0.0003726	0.0429038
2017-01-11	0.0026917	0.0028261	0.0007449	-0.0530961
2017-01-12	-0.0015456	-0.0025099	-0.0010234	0.9404691
2017-01-13	0.0034219	0.0022955	0.0029803	0.0220018
2017-01-17	-0.0030042	-0.0035235	-0.0015787	-0.0012128
2017-01-18	0.0020360	0.0022099	0.0019532	0.0428051

Compute the returns from the closing prices

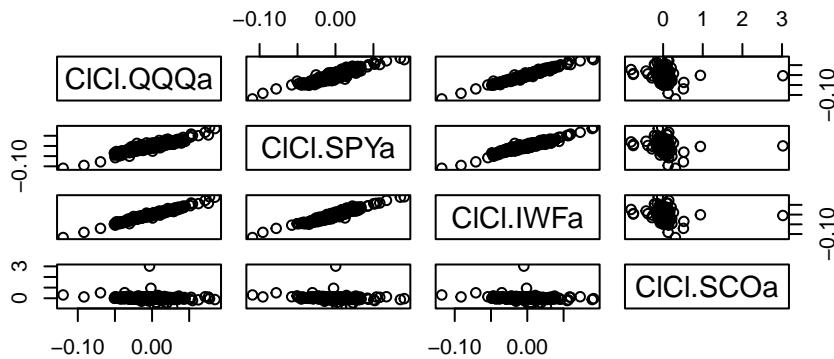


Figure 13: We can see a strong linear correlation here in 3 of the stocks QQQ, SPY and IWF. SCO seems to not have a strong correlation with the other stocks

### Volatility of the ETFs

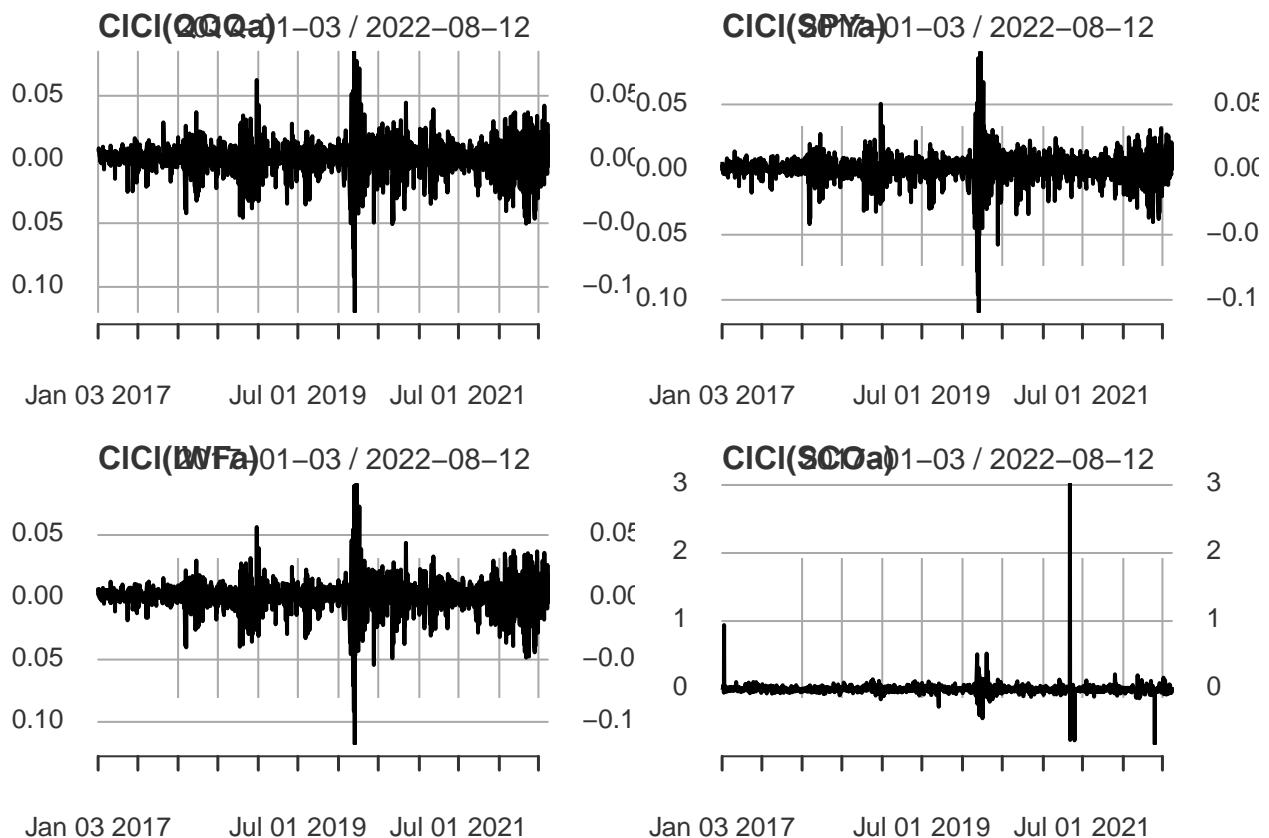


Figure 14: Volatility of the ETFs across the 5 year period.

## Sample a random return from the empirical joint distribution

We simulated a random day's return.

```
##          C1C1.QQQa  C1C1.SPYa  C1C1.IWFa  C1C1.SCOa
## 2020-11-16  0.00780263  0.01248255  0.00525364 -0.0415355
```

## SIMULATION 1: LOW RISK PORTFOLIO

- INITIAL INVESTMENT IS \$ 100000.
- Average return of investment after 20 days = \$ 101190.
- 5% value at Risk for safe portfolio = \$ 12139.7.

Now simulate many different possible futures, repeating the above block thousands of times

```
##          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## result.1  100025.03  99877.23  100193.04  100374.26  99625.71  96487.93  97413.44
## result.2   97645.18  97019.54  96239.56  97457.63  96709.73  97313.76  97977.01
## result.3   99517.79  99492.94  101159.87  100681.74  101423.88  103183.56  103096.60
## result.4  101442.00  100935.60  103448.92  102039.49  103123.97  104118.65  104762.17
## result.5   99474.78  100467.10  101100.43  100714.50  101333.30  100132.13  98823.24
## result.6  100263.80  100834.35  100448.19  98329.78  98085.05  97114.86  98392.72
## result.7  100328.18  96501.36  96639.18  95979.78  97169.25  96641.63  96628.56
## result.8   98975.37  100279.24  100088.42  99671.95  99929.20  99933.75  99888.97
## result.9   99905.17  100956.47  100581.08  101405.10  103603.35  102064.87  100580.00
## result.10  99016.41  98472.57  97838.74  98517.02  103810.87  103636.66  104349.01
##          [,8]      [,9]      [,10]     [,11]     [,12]     [,13]     [,14]
## result.1  100497.09  99921.46  99940.86  99408.28  98408.91  98566.87  102087.43
## result.2   98229.31  97115.23  96891.40  96291.05  93609.72  93050.87  93353.84
## result.3  105078.23  104508.88  104390.85  105181.15  105231.31  105707.43  105131.59
## result.4  103812.60  104813.87  105877.52  105126.08  104543.94  105175.52  104733.71
## result.5  101026.30  101897.69  101160.75  100678.54  100731.01  102682.47  102651.92
## result.6   98707.96  99895.90  99181.88  97716.62  97625.80  97955.81  98290.08
## result.7   96875.78  96959.73  97005.34  96783.23  96795.33  97669.40  95895.67
## result.8   99888.89  102162.06  101998.81  101915.25  102177.09  102091.57  104594.90
## result.9  100204.84  99186.45  99399.80  100087.00  99920.35  101661.35  101132.93
## result.10  102939.02  102159.35  100771.05  100440.34  99674.24  101909.53  99559.33
##          [,15]     [,16]     [,17]     [,18]     [,19]     [,20]
## result.1   92692.72  92000.06  91914.92  90919.68  92442.49  93350.74
## result.2   93651.97  94333.45  93612.61  95906.43  98141.48  97279.79
## result.3  104886.14  96130.76  95229.86  95265.98  95383.47  96780.00
## result.4  105039.36  107250.54  107578.46  109943.38  109773.35  111440.80
## result.5  103268.40  103318.04  103244.54  102805.14  101819.84  100588.24
## result.6  101722.34  101197.23  100690.52  100525.19  99684.61  99044.00
## result.7   95383.83  95273.85  94962.64  95355.06  95114.06  93967.50
## result.8  104077.90  103740.78  104200.30  103394.30  102548.63  102295.37
## result.9  101608.46  101363.34  101078.19  101893.09  101444.90  98940.62
## result.10  100922.43  99985.79  100372.86  100986.85  101472.71  100511.51
```

### Histogram of sim1[, n\_days]

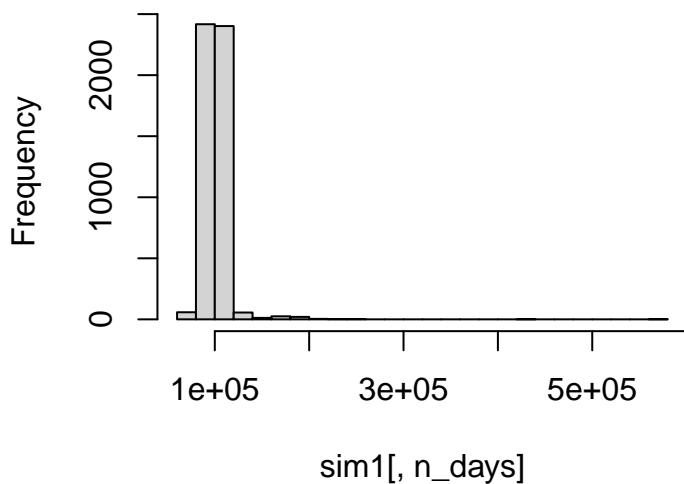


Figure 15: Histogram and density plots of simulation 1. We see spikes at 100000

### density.default(x = sim1[, n\_days])

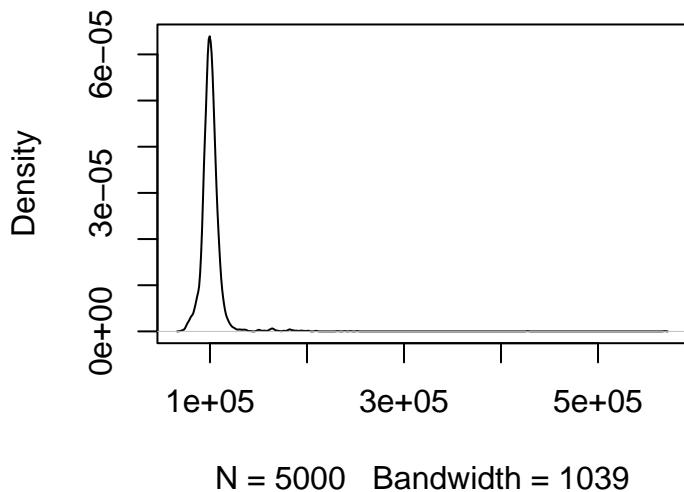


Figure 16: Histogram and density plots of simulation 1. We see spikes at 100000

## Profit and Loss

- Average return of investment after 20 days = \$101099.5
- Average profit/loss after 20 days \$1099.513

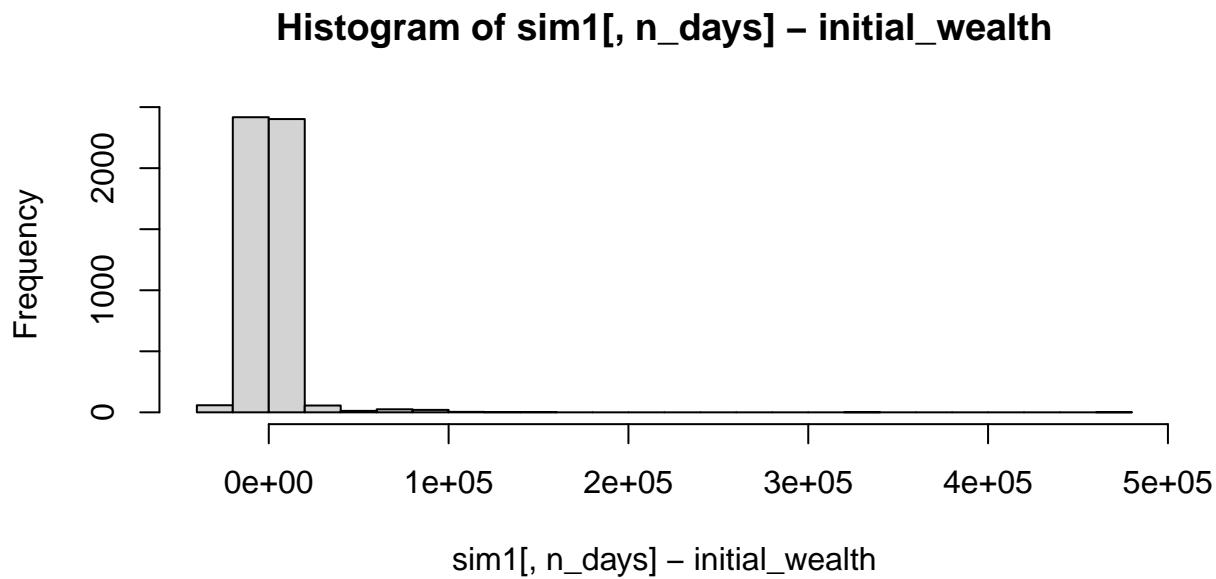


Figure 17: Histogram of returns of simulation 1.

5% Value at Risk for the first simulation = -12293.44

## SIMULATION 2: HIGH RISK PORTFOLIO

- Average return of investment after 20 days = \$101576.2
- 5% Value at Risk for safe portfolio = \$7849.9327

```
##          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## result.1 99907.76 100563.85 99311.09 99145.26 98258.92 97870.60 98096.91
## result.2 96421.68 95709.17 96480.34 97416.92 97205.12 97304.13 96559.88
## result.3 98847.61 98297.00 98897.86 99083.52 97876.25 97580.36 97554.75
## result.4 99555.86 99599.53 99222.58 100224.08 100049.35 100158.74 101383.49
## result.5 99291.32 97720.32 97720.23 96587.07 95845.40 96294.77 96396.97
## result.6 94178.60 94695.30 93407.15 93004.70 91907.62 93968.69 96135.22
## result.7 99823.12 101757.13 103057.75 103200.95 103786.53 104699.39 107407.01
## result.8 100026.06 100959.34 101280.08 101432.75 101672.86 103063.27 103388.54
## result.9 100325.59 100201.62 96635.85 96335.01 96193.59 98466.79 97886.42
## result.10 100079.63 100026.30 101068.65 101388.47 100124.58 101676.39 101367.76
##          [,8]      [,9]      [,10]     [,11]     [,12]     [,13]     [,14]
## result.1 97997.72 99506.52 99447.35 99147.45 99513.31 99233.34 98532.11
## result.2 96465.46 97009.71 97688.41 97726.21 98408.27 98850.31 98790.54
## result.3 98417.63 97376.23 97912.89 98587.93 99073.43 99380.47 99024.88
## result.4 100542.43 101598.71 100930.16 100635.26 100127.31 100339.57 102228.89
## result.5 97325.09 98023.56 98587.17 100522.08 100993.53 99316.36 99897.42
## result.6 96179.49 99046.87 99574.51 98005.82 97628.42 97487.29 97221.83
## result.7 106637.05 108026.19 108553.35 110403.42 110042.80 111043.26 109897.60
## result.8 104223.51 104399.94 104441.77 104545.24 103631.42 105103.41 106268.47
## result.9 98203.62 97103.66 96590.26 96045.98 94795.48 93839.58 93785.12
## result.10 100254.75 99451.33 100918.87 102087.72 101846.09 99467.33 100171.27
```

```

##          [,15]      [,16]      [,17]      [,18]      [,19]      [,20]
## result.1 98440.52 98742.37 98265.94 95761.65 96660.49 97118.14
## result.2 98513.06 97978.35 97369.96 97633.73 97007.64 97944.39
## result.3 100111.40 99759.35 99343.79 98541.59 98933.74 99881.66
## result.4 103643.75 106359.61 108125.64 109049.77 109310.45 109202.30
## result.5 100465.25 101331.88 98242.46 96647.14 96621.52 98259.03
## result.6 96937.26 103138.52 103176.94 103231.04 104727.83 102311.15
## result.7 109781.52 110642.35 108996.90 110233.33 110079.54 111536.82
## result.8 105889.01 106032.97 103113.24 103446.98 102708.44 103268.51
## result.9 93603.61 94289.91 93446.53 93208.80 93421.20 94320.68
## result.10 101232.21 100801.23 100884.21 101013.67 100608.26 101997.31

```

### Histogram of sim2[, n\_days]

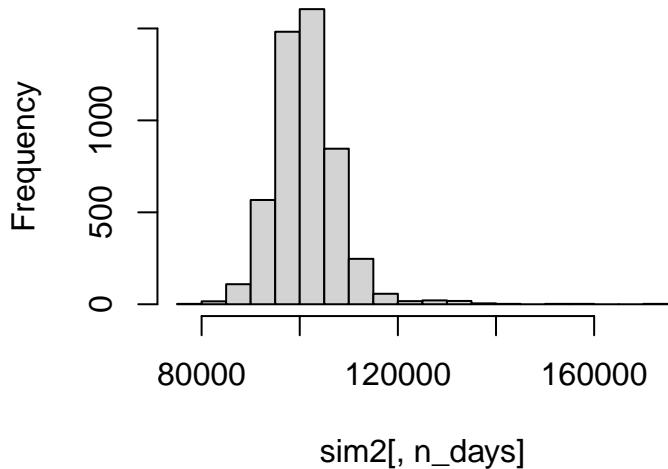


Figure 18: Histogram and density plots of simulation 2. We see spikes at 100000

**density.default(x = sim2[, n\_days])**

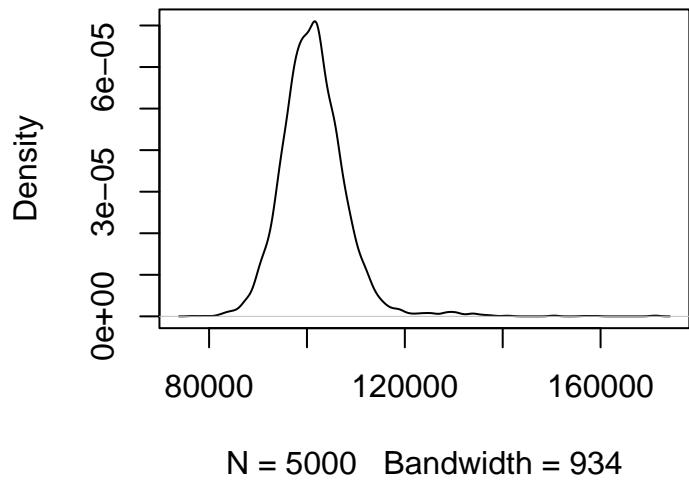


Figure 19: Histogram and density plots of simulation 2. We see spikes at 100000

### Profit and Loss

**Histogram of sim2[, n\_days] – initial\_wealth**

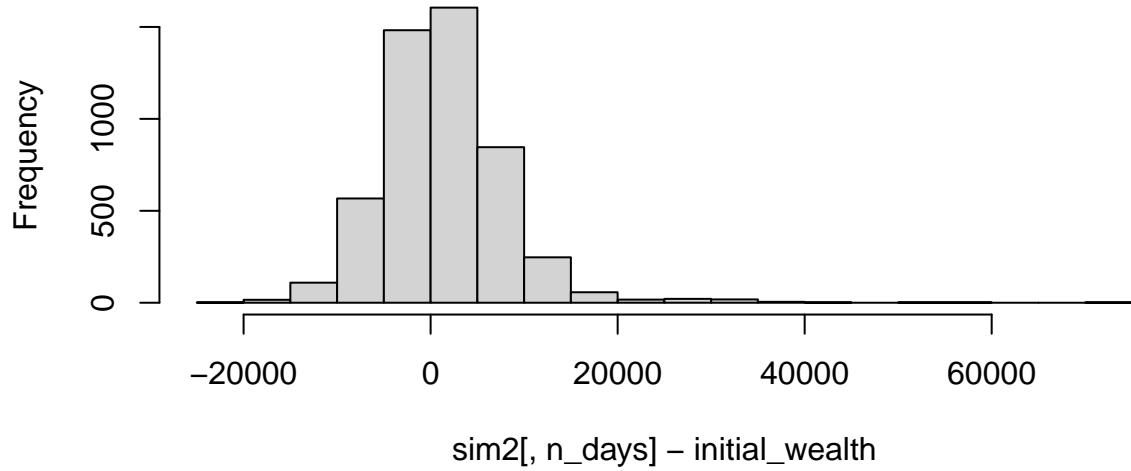


Figure 20: Histogram of returns of simulation 2.

5% Value at Risk for the second simulation = -8281.054

## Summary

For the safe portfolio, we are observing lower return of investment and lower 5% VaR.

As the portfolio risk increased, we are able to witness the increase in returns and an increase in VaR value as expected.

References: <https://www.bankrate.com/investing/best-etfs/> <https://etfdb.com/compare/lowest-ytd-returns/>

## 6. Clustering and PCA

### Data Summary

```
##   fixed.acidity    volatile.acidity    citric.acid    residual.sugar
##   Min. : 3.800    Min. :0.0800    Min. :0.0000    Min. : 0.600
##   1st Qu.: 6.400    1st Qu.:0.2300    1st Qu.:0.2500    1st Qu.: 1.800
##   Median : 7.000    Median :0.2900    Median :0.3100    Median : 3.000
##   Mean   : 7.215    Mean   :0.3397    Mean   :0.3186    Mean   : 5.443
##   3rd Qu.: 7.700    3rd Qu.:0.4000    3rd Qu.:0.3900    3rd Qu.: 8.100
##   Max.   :15.900    Max.   :1.5800    Max.   :1.6600    Max.   :65.800
##   chlorides      free.sulfur.dioxide total.sulfur.dioxide    density
##   Min. :0.00900    Min. : 1.00    Min. : 6.0     Min. :0.9871
##   1st Qu.:0.03800    1st Qu.:17.00    1st Qu.:77.0    1st Qu.:0.9923
##   Median :0.04700    Median :29.00    Median :118.0    Median :0.9949
##   Mean   :0.05603    Mean   :30.53    Mean   :115.7    Mean   :0.9947
##   3rd Qu.:0.06500    3rd Qu.:41.00    3rd Qu.:156.0    3rd Qu.:0.9970
##   Max.   :0.61100    Max.   :289.00    Max.   :440.0    Max.   :1.0390
##   pH      sulphates      alcohol      quality
##   Min. :2.720    Min. :0.2200    Min. : 8.00    Min. :3.000
##   1st Qu.:3.110    1st Qu.:0.4300    1st Qu.: 9.50    1st Qu.:5.000
##   Median :3.210    Median :0.5100    Median :10.30    Median :6.000
##   Mean   :3.219    Mean   :0.5313    Mean   :10.49    Mean   :5.818
##   3rd Qu.:3.320    3rd Qu.:0.6000    3rd Qu.:11.30    3rd Qu.:6.000
##   Max.   :4.010    Max.   :2.0000    Max.   :14.90    Max.   :9.000
##   color
##   Length:6497
##   Class :character
##   Mode  :character
## 
## 
```

### PCA

We used prcomp that uses the singular value decomposition (SVD).

```
## Importance of components:
##                               PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.72330
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.04756
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732 0.85253 0.90009
##                               PC8     PC9     PC10    PC11
## Standard deviation     0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion  0.94568 0.97632 0.9970 1.00000
```

## Interpreting the loadings

Summary The loadings are as follows

```
##          Properties      PC1      PC2      PC3      PC4
## 1    fixed.acidity -0.23879890  0.33635454 -0.43430130  0.16434621
## 2   volatile.acidity -0.38075750  0.11754972  0.30725942  0.21278489
## 3     citric.acid   0.15238844  0.18329940 -0.59056967 -0.26430031
## 4  residual.sugar   0.34591993  0.32991418  0.16468843  0.16744301
## 5     chlorides    -0.29011259  0.31525799  0.01667910 -0.24474386
## 6 free.sulfur.dioxide  0.43091401  0.07193260  0.13422395 -0.35727894
## 7 total.sulfur.dioxide  0.48741806  0.08726628  0.10746230 -0.20842014
## 8       density    -0.04493664  0.58403734  0.17560555  0.07272496
## 9         pH     -0.21868644 -0.15586900  0.45532412 -0.41455110
## 10    sulphates   -0.29413517  0.19171577 -0.07004248 -0.64053571
## 11     alcohol   -0.10643712 -0.46505769 -0.26110053 -0.10680270
##          PC5      PC6      PC7      PC8      PC9      PC10
## 1 -0.1474804 -0.20455371 -0.28307944  0.401235645  0.3440567 -0.281267685
## 2  0.1514560 -0.49214307 -0.38915976 -0.087435088 -0.4969327  0.152176731
## 3 -0.1553487  0.22763380 -0.38128504 -0.293412336 -0.4026887  0.234463340
## 4 -0.3533619 -0.23347775  0.21797554 -0.524872935  0.1080032 -0.001372773
## 5  0.6143911  0.16097639 -0.04606816 -0.471516850  0.2964437 -0.196630217
## 6  0.2235323 -0.34005140 -0.29936325  0.207807585  0.3666563  0.480243340
## 7  0.1581336 -0.15127722 -0.13891032  0.128621319 -0.3206955 -0.713663486
## 8 -0.3065613  0.01874307 -0.04675897  0.004831136  0.1128800 -0.003908289
## 9 -0.4533764  0.29657890 -0.41890702 -0.028643277  0.1278367 -0.141310977
## 10 -0.1365769 -0.29692579  0.52534311  0.165818022 -0.2077642  0.045959499
## 11 -0.1888920 -0.51837780 -0.10410343 -0.399233887  0.2518903 -0.205053085
##          PC11
## 1 -0.3346792663
## 2 -0.0847718098
## 3  0.0011089514
## 4 -0.4497650778
## 5 -0.0434375867
## 6  0.0002125351
## 7  0.0626848131
## 8  0.7151620723
## 9 -0.2063605036
## 10 -0.0772024671
## 11  0.3357018784
```

PC1 is always the axis that explains more variability among the samples included in the test. PC2 is the second axis.

Table 5: Weights on the components

	PC1	PC2
fixed.acidity	-0.24	0.34
volatile.acidity	-0.38	0.12
citric.acid	0.15	0.18
residual.sugar	0.35	0.33
chlorides	-0.29	0.32
free.sulfur.dioxide	0.43	0.07
total.sulfur.dioxide	0.49	0.09
density	-0.04	0.58

	PC1	PC2
pH	-0.22	-0.16
sulphates	-0.29	0.19
alcohol	-0.11	-0.47

Table 6: Summary of PC1

Properties	PC1
total.sulfur.dioxide	0.49
free.sulfur.dioxide	0.43
residual.sugar	0.35
citric.acid	0.15
density	-0.04
alcohol	-0.11
pH	-0.22
fixed.acidity	-0.24
chlorides	-0.29
sulphates	-0.29
volatile.acidity	-0.38

PC1 seems to give more positive loadings to sulfur dioxide and residual sugar while chlorides, sulphates & acidity have negative loadings

Table 7: Summary of PC2

Properties	PC2
density	0.58
fixed.acidity	0.34
residual.sugar	0.33
chlorides	0.32
sulphates	0.19
citric.acid	0.18
volatile.acidity	0.12
total.sulfur.dioxide	0.09
free.sulfur.dioxide	0.07
pH	-0.16
alcohol	-0.47

PC2 gives more positive loading to density and high negative to alcohol content

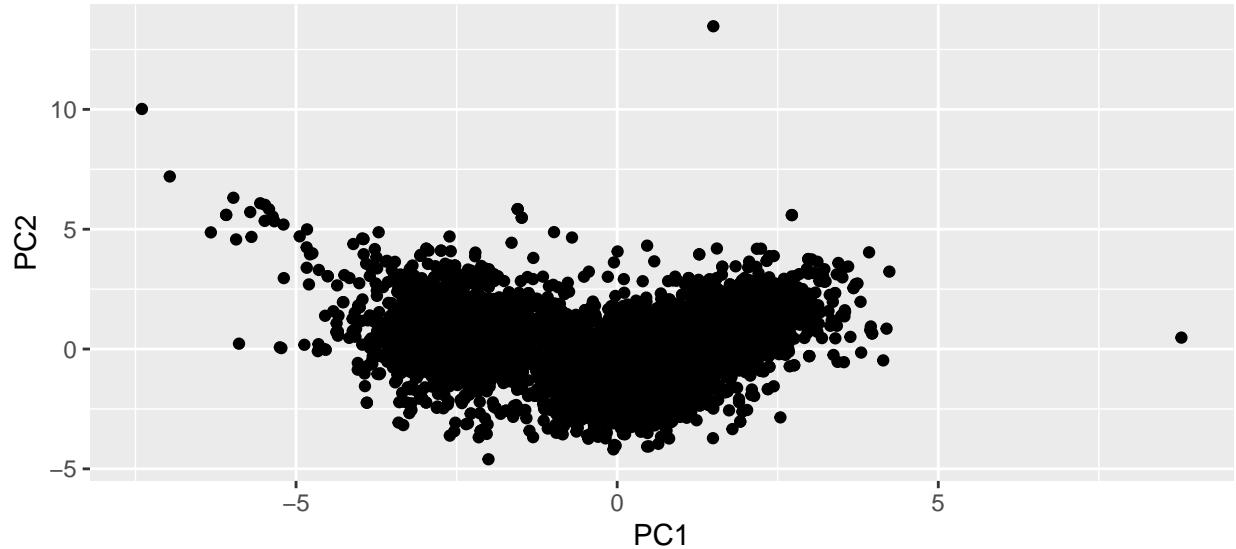


Figure 21: PC space, i.e. the space of summary variables we've created

- We can see that there are two clusters emerging. One to the left having more negative PC1 and another to the right having more positive PC1.
- As we do not have access to the wine color, we cannot say exactly which of these two clusters correspond to red and White wine.
- However, the difference in the loadings for each feature by the principal components implies that these two clusters are distinct.

### Categorize quality

Here we predict and categorize quality.

Table 8: Quality

Var1	Freq
High Quality	4113
Low Quality	2384

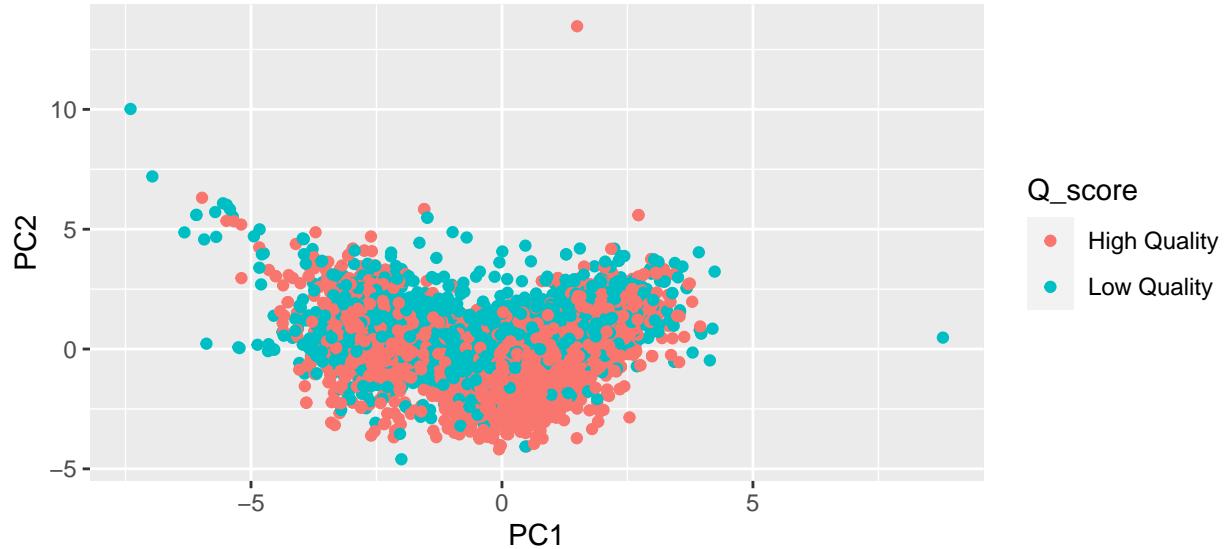


Figure 22: It doesn't look like the principal components can do a good job of separating high and low quality wines

## Clusters

Table 9: Cluster center 1

	x
fixed.acidity	6.8516790
volatile.acidity	0.2745838
citric.acid	0.3352493
residual.sugar	6.3940255
chlorides	0.0451042
free.sulfur.dioxide	35.5215286
total.sulfur.dioxide	138.4584878
density	0.9940049
pH	3.1876246
sulphates	0.4888051
alcohol	10.5223589

Table 10: Cluster center 2

	x
fixed.acidity	8.2895922
volatile.acidity	0.5319416
citric.acid	0.2695435
residual.sugar	2.6342666
chlorides	0.0883238
free.sulfur.dioxide	15.7647596
total.sulfur.dioxide	48.6396835
density	0.9967404
pH	3.3097200

	x
sulphates	0.6567194
alcohol	10.4015216

## Observations

- More sulfur dioxide and sugar compared to cluster 2.
- More acidity, chlorides and sulphates compared to cluster 1.

## How close the clusters are to each other

Table 11: Cluster center-1 to cluster center-2

	x
volatile.acidity	-1.5631878
chlorides	-1.2336600
sulphates	-1.1284118
fixed.acidity	-1.1091297
density	-0.9122427
pH	-0.7593601
alcohol	0.1013131
citric.acid	0.4521520
residual.sugar	0.7902299
free.sulfur.dioxide	1.1130951
total.sulfur.dioxide	1.5890987

This shows that the cluster centers are quite close to each other.

## Table clusters

Table 12: Confusion Matrix

red	white
24	4830
1575	68

Table 13: First 10 rows of PCs for wine data

dist1	dist2
107.40131	15.47603
72.34191	20.57183
87.02246	5.48984
80.86079	11.83767
107.40131	15.47603
101.11543	9.21505

## Visualization

### Contrasting features

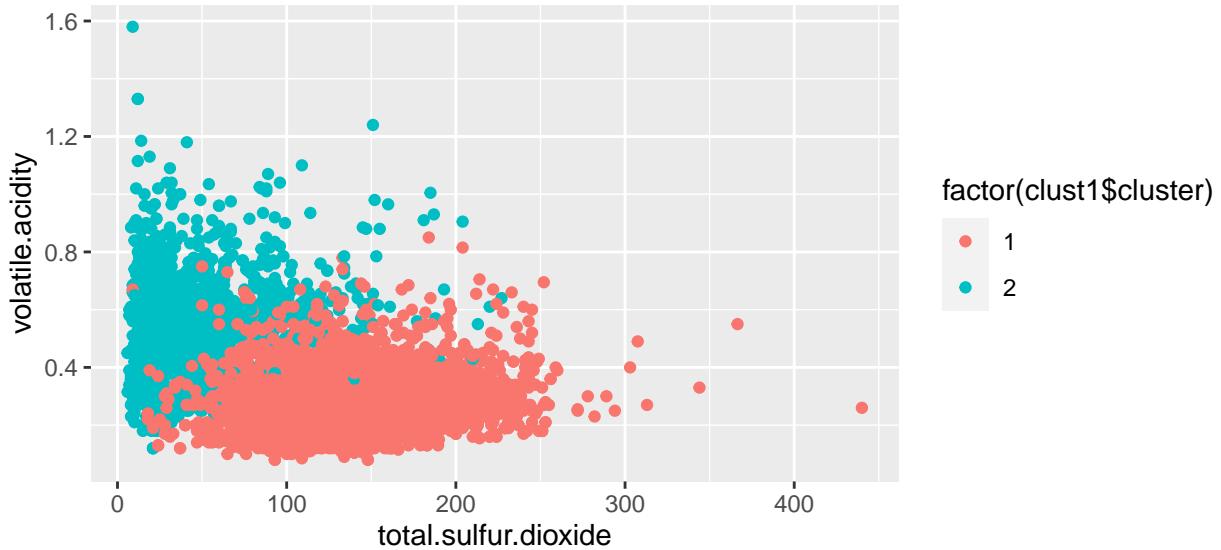


Figure 23: Using the two most contrasting features for the 2 clusters as x and y to visualize

### Plot for High and Low Quality Wine

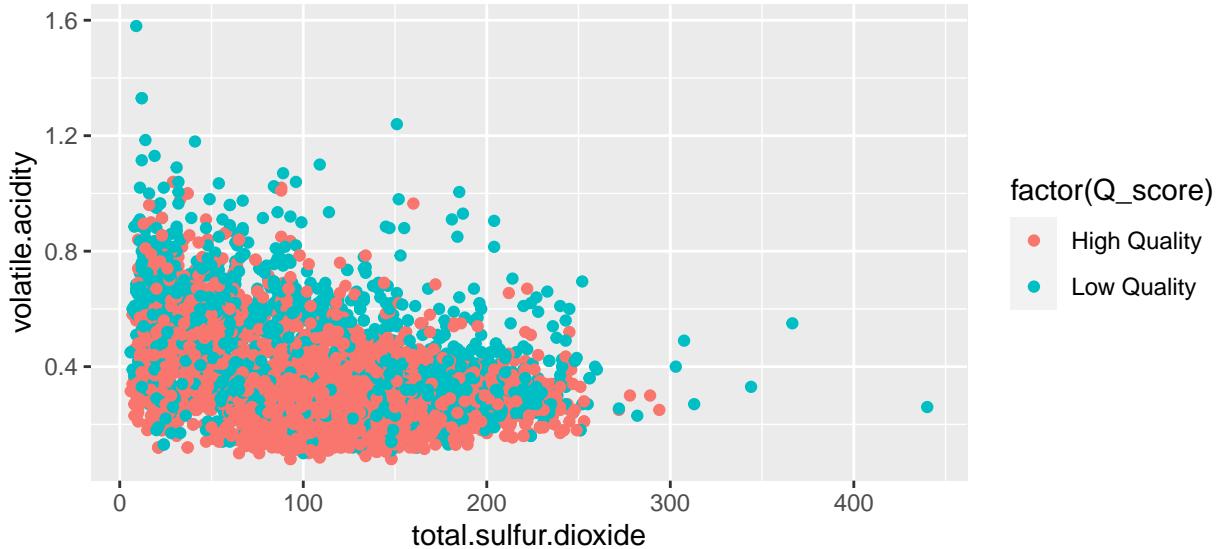


Figure 24: We can see that the features are not very discriminative of quality.

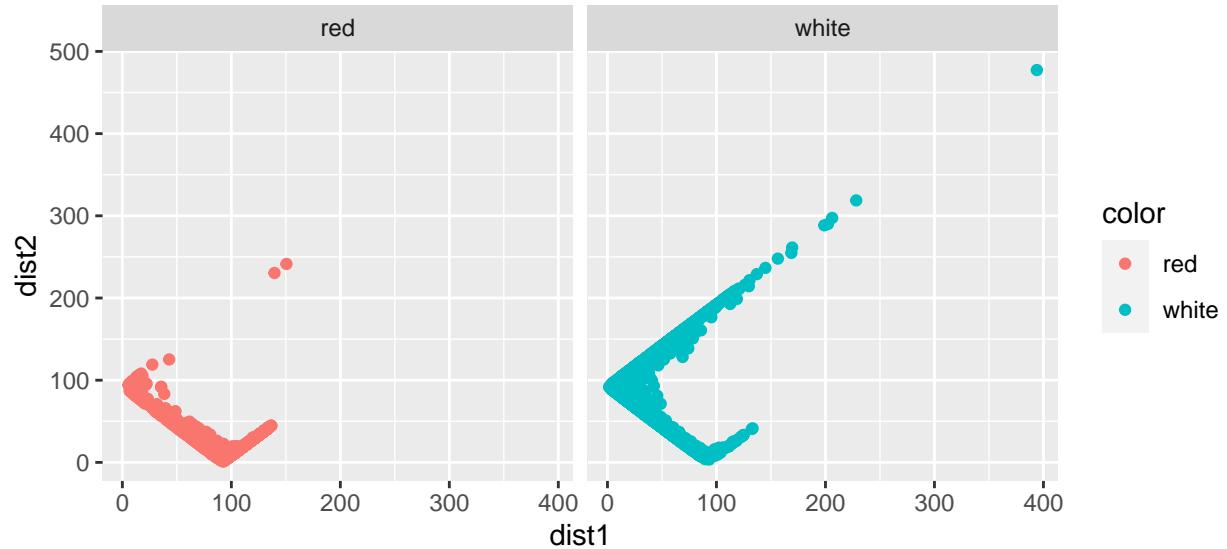


Figure 25: Interaction of red and white wine variables individually in the plot

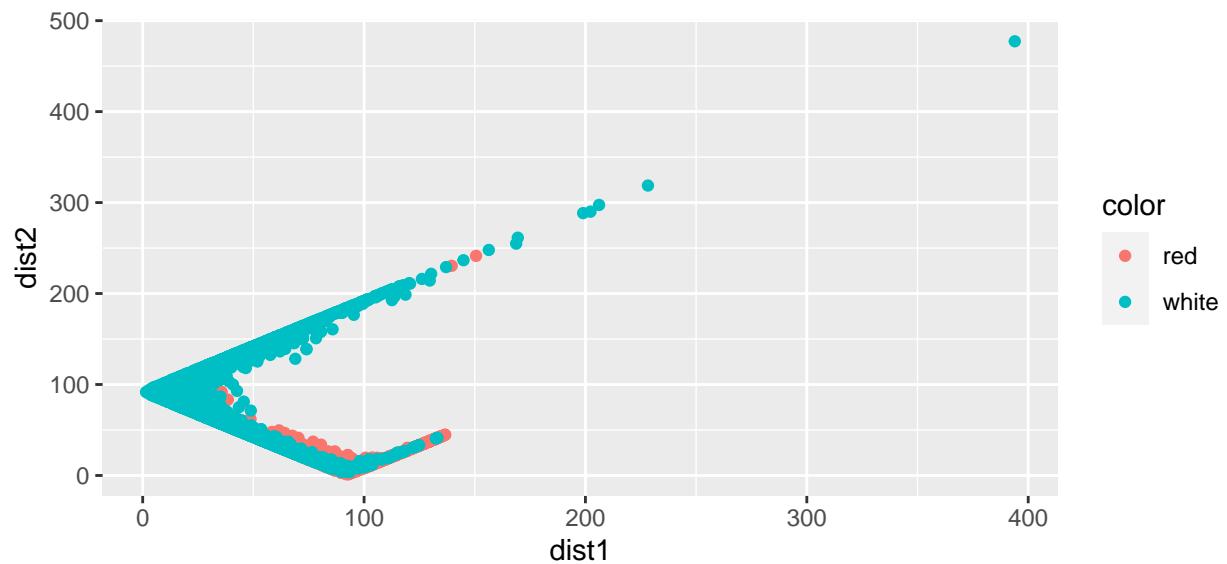


Figure 26: Interaction of red and white wine variables in the plot

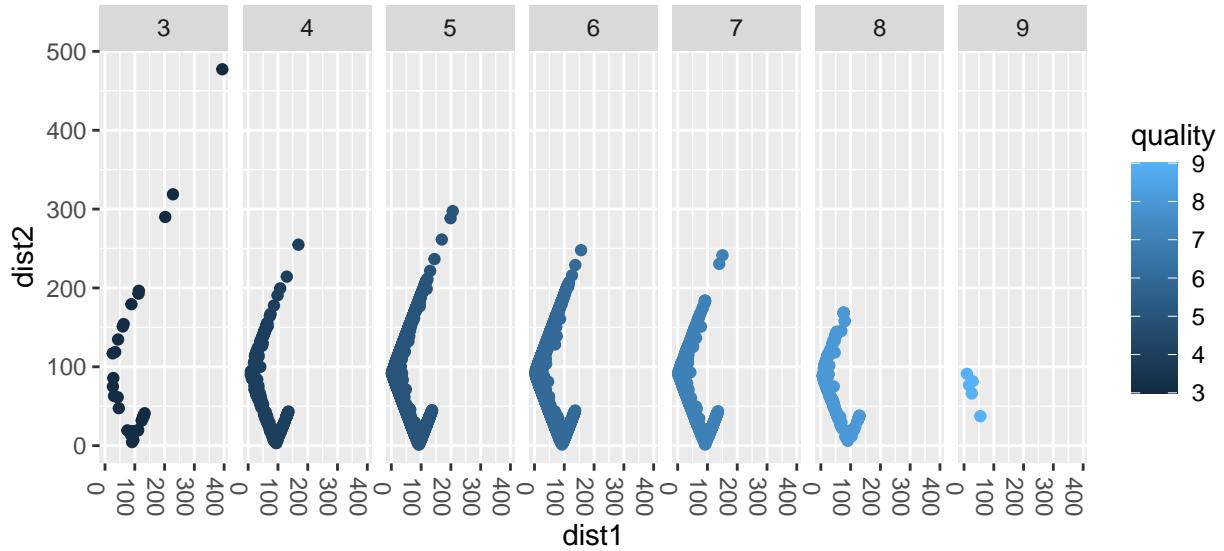


Figure 27: Interaction of red and white wine variables wrt quality parameter in the plot.

## Summary

- PCA is the best dimensionality reduction technique makes more sense to us for this data.
- PCA Counteracts the issues of high dimensional data.
- PCA improves performance at a very low cost of model accuracy.

## 7. Market segmentation

### Correlation Plot

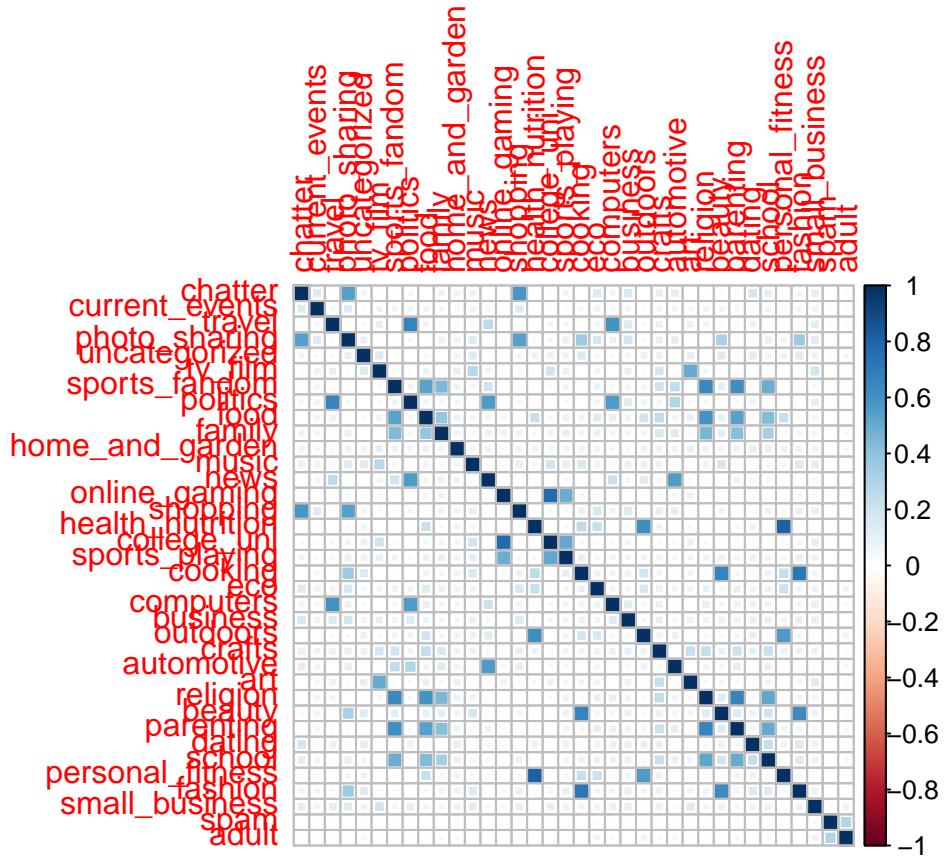


Figure 28: The plot is showing us that college uni and online gaming are correlated. We also see that health nutrition and personal fitness are correlated as well.

We can use the principal component analysis to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed.

We've centered and scaled the data and extracted them which will be the attributes.

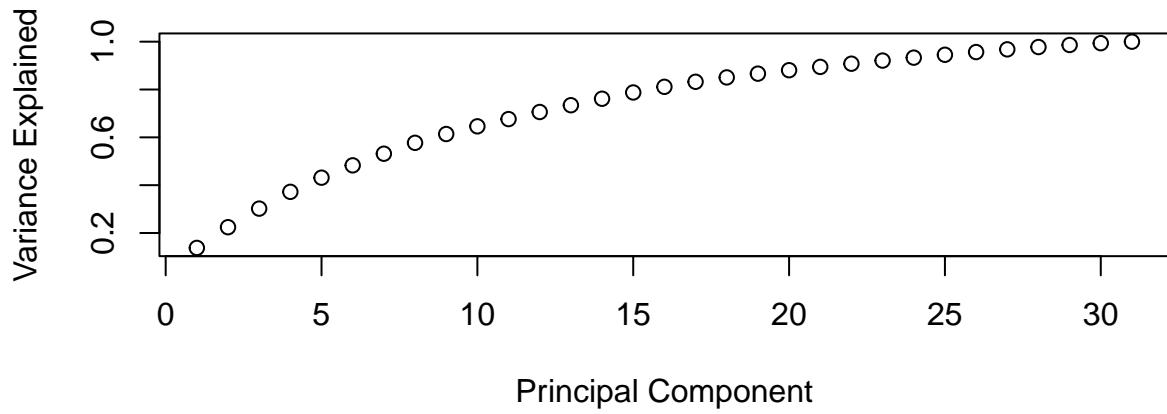


Figure 29: Looking at the 10th principal component we can see that roughly 64.6 percent of the fraction of variance is explained. Using the Kaiser Criterion we can drop all Eigen values under 1.0.

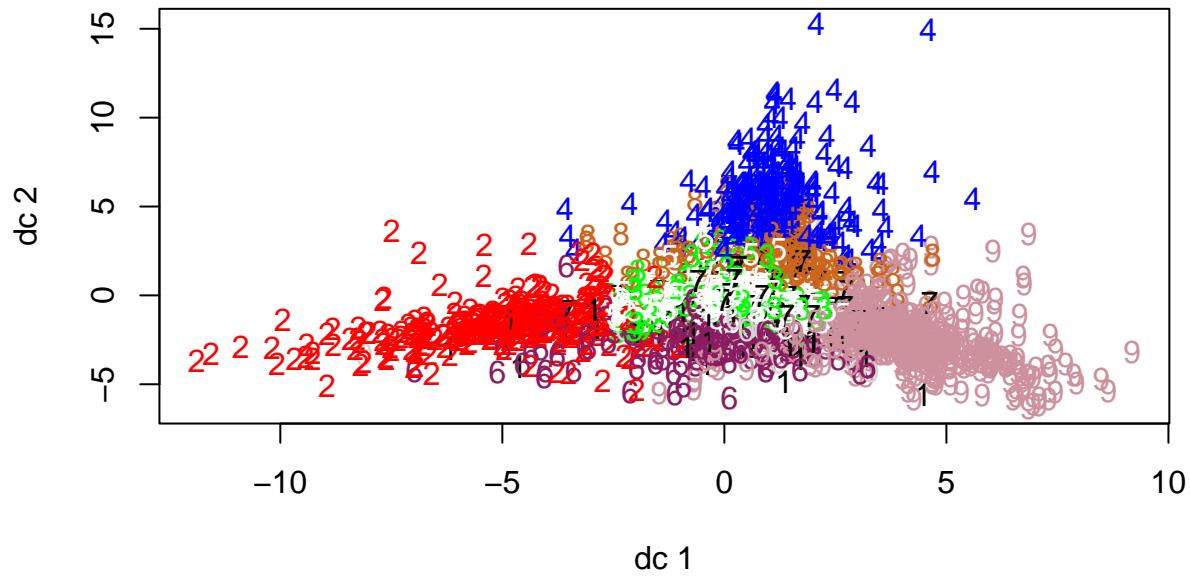


Figure 30: Visualization of the clusters. Using the elbow method we've determine the number of clusters. We've decided to move forward with 10 clusters.

### Summary of the principal components

##	V1	V2	V3	V4
----	----	----	----	----

```

##  Min.   :0.5477   Min.   : 0.3408   Min.   :0.4142   Min.   : 0.4153
##  1st Qu.:0.8090   1st Qu.: 0.5824   1st Qu.:0.5885   1st Qu.: 0.6568
##  Median :1.0804   Median : 0.8380   Median :0.7673   Median : 0.8249
##  Mean   :1.3856   Mean   : 1.5891   Mean   :0.9776   Mean   : 1.6967
##  3rd Qu.:1.3894   3rd Qu.: 1.2710   3rd Qu.:1.1911   3rd Qu.: 1.3630
##  Max.   :9.2814    Max.   :11.0950   Max.   :3.3613   Max.   :11.1836
##          V5           V6           V7           V8
##  Min.   :0.1760   Min.   : 0.4351   Min.   :0.2504   Min.   :0.2432
##  1st Qu.:0.3527   1st Qu.: 0.7732   1st Qu.:0.5915   1st Qu.:0.5462
##  Median :0.4488   Median : 0.9052   Median :0.7804   Median :0.8671
##  Mean   :0.5359   Mean   : 1.5799   Mean   :1.1924   Mean   :1.3236
##  3rd Qu.:0.7067   3rd Qu.: 1.3495   3rd Qu.:1.1384   3rd Qu.:1.1655
##  Max.   :1.2408   Max.   :11.6825   Max.   :6.8858   Max.   :6.7725
##          V9           V10
##  Min.   :0.3933   Min.   : 0.4982
##  1st Qu.:0.7193   1st Qu.: 0.6697
##  Median :1.0175   Median : 0.9711
##  Mean   :1.5999   Mean   :1.4437
##  3rd Qu.:1.5344   3rd Qu.: 1.5758
##  Max.   :6.1199   Max.   :7.1336

```

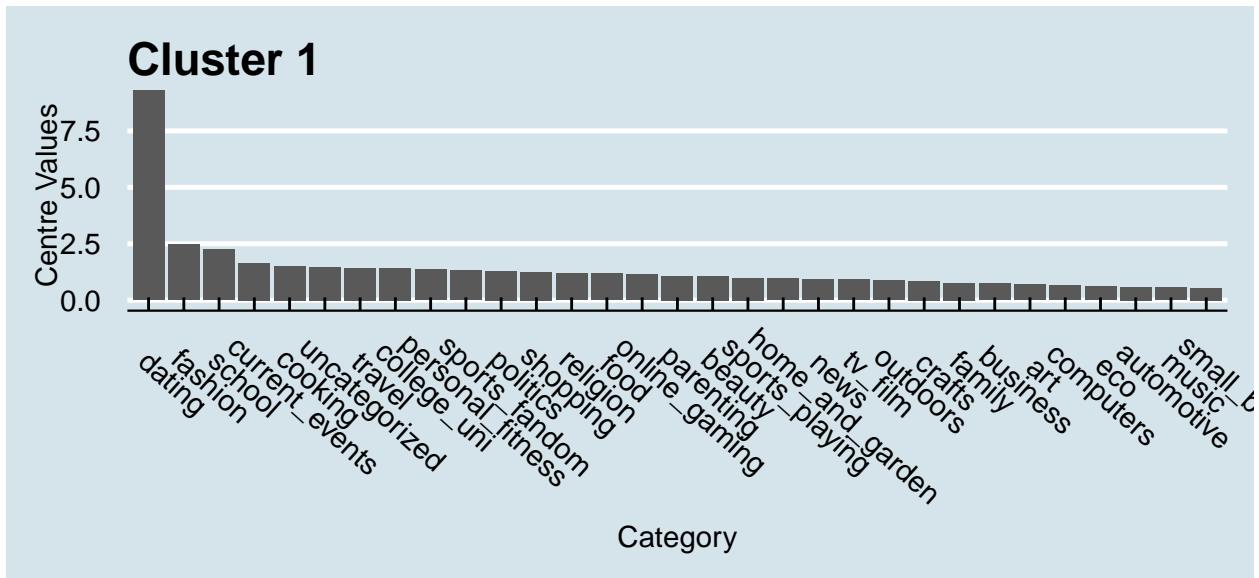


Figure 31: Cluster 1 primarily composed with people who have interest in current events, traveling and politics.

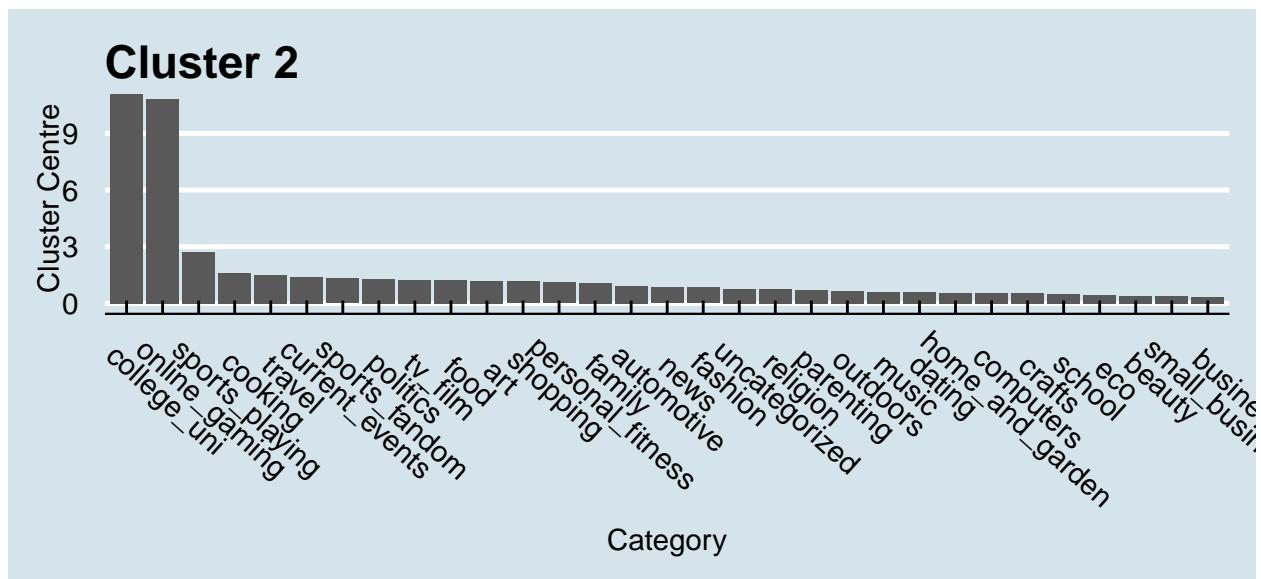


Figure 32: Cluster 2 primarily composed with people who have interest in shopping, current events, college uni.

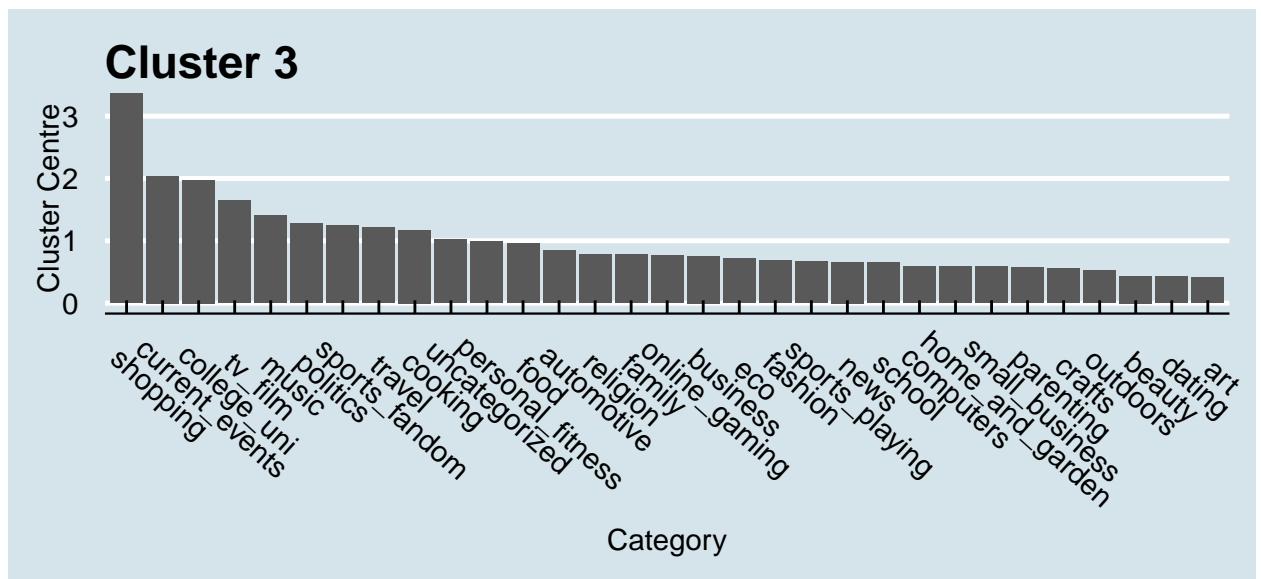


Figure 33: Cluster 3 emphasized people who have interest in dating then followed by fashion.

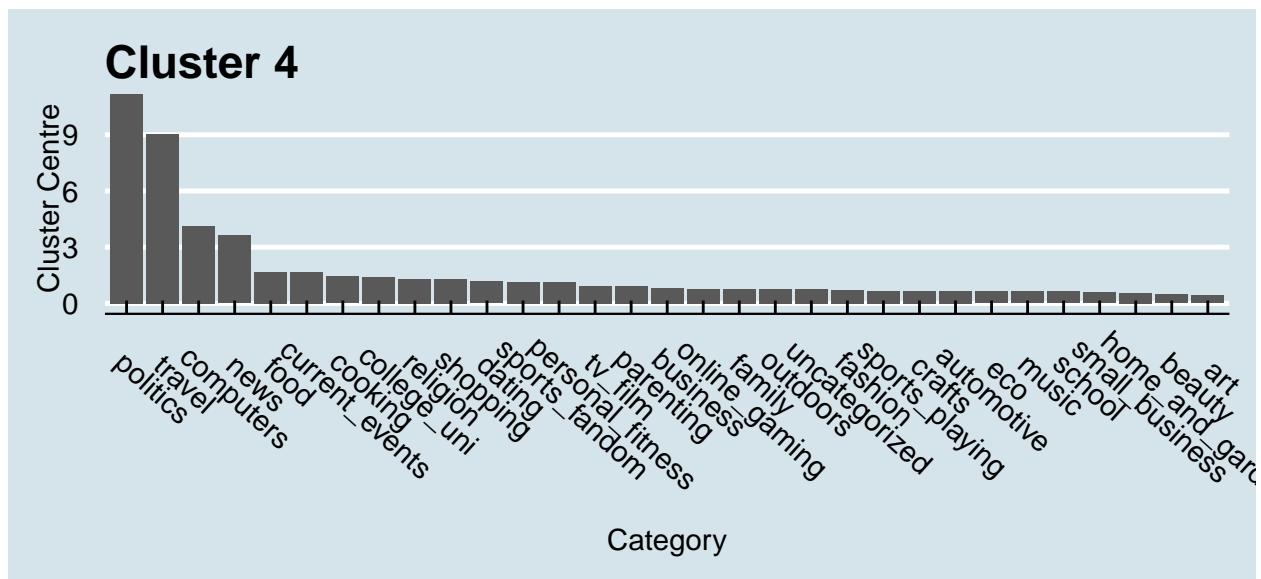


Figure 34: Cluster 4 primarily composed of people who have interest in sports fandom, religion, and parenting.

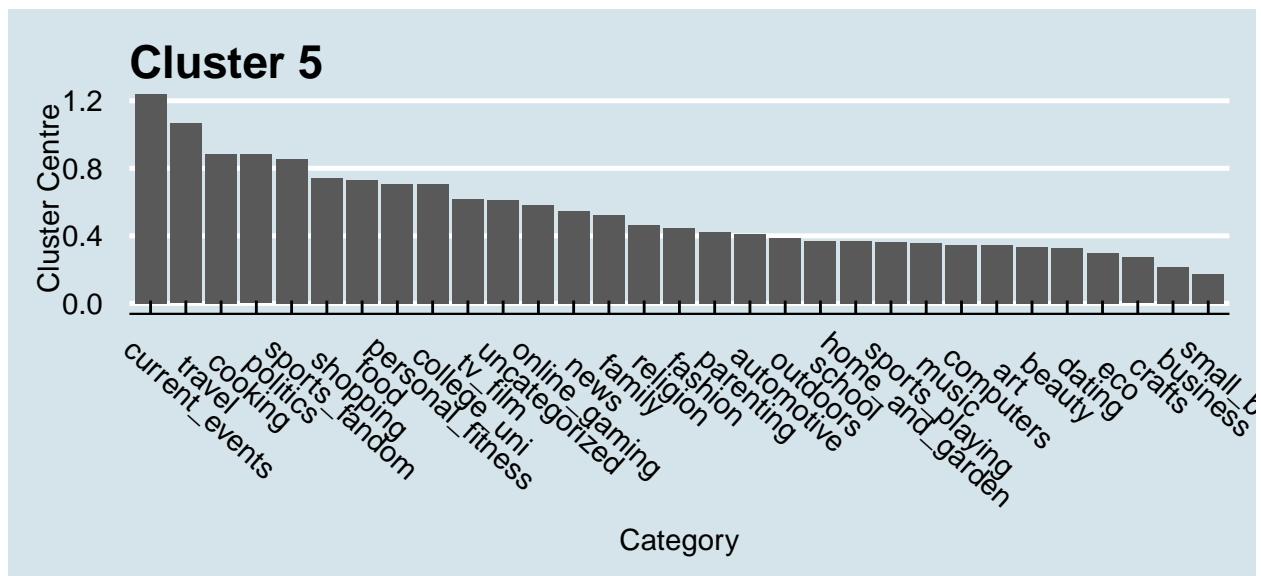


Figure 35: Cluster 5 primarily composed of people who have interest in cooking, fashion and beauty.

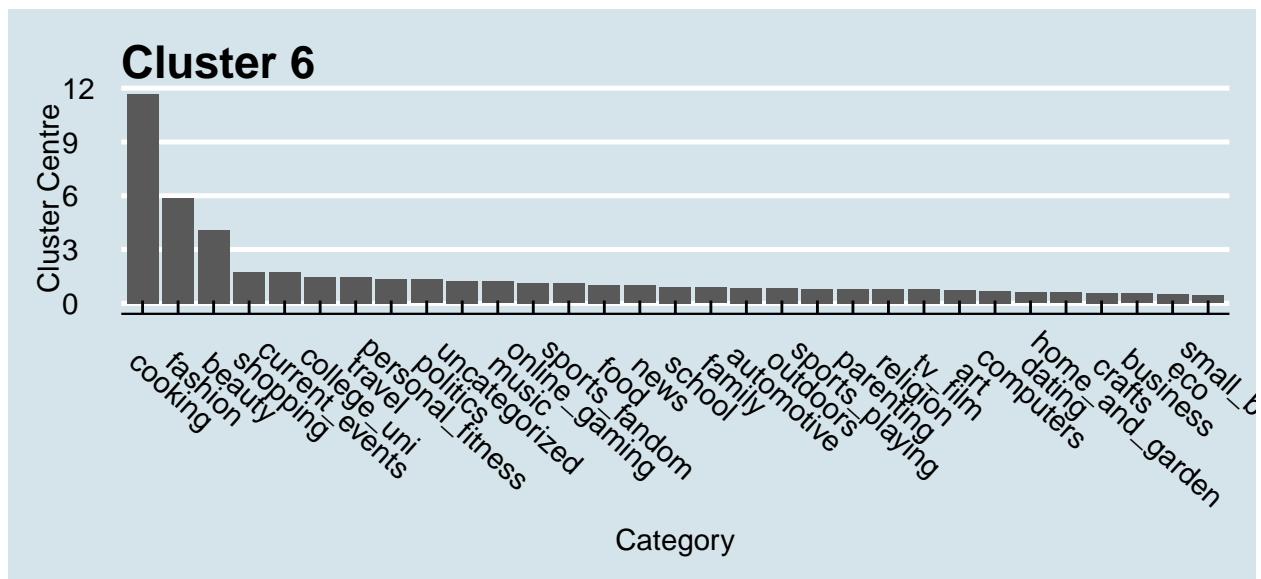


Figure 36: Cluster 6 primarily composed of people who have interest in college unit, online gaming and sports playing.

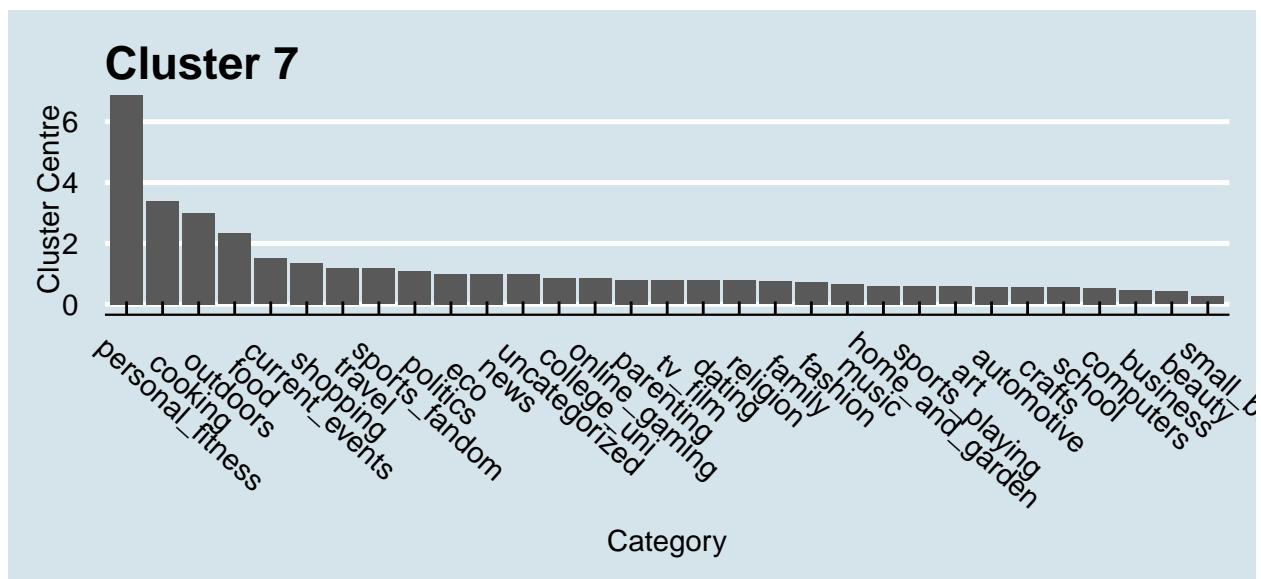


Figure 37: Cluster 7 primarily composed of people who have interest in art, tv film, and current events.

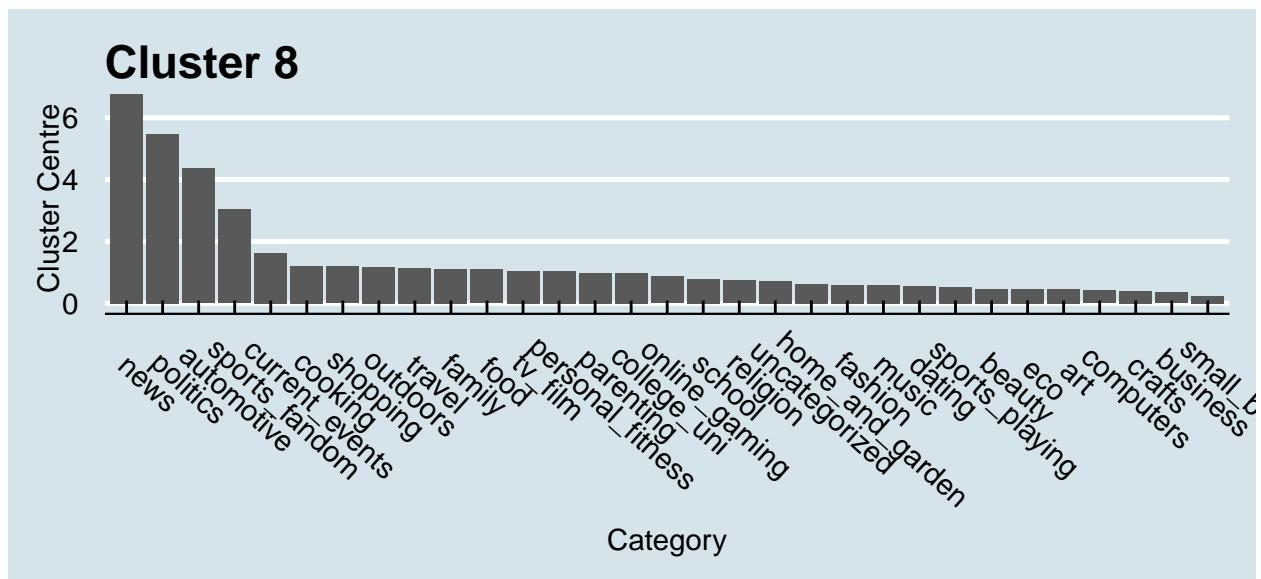


Figure 38: Cluster 8 primarily composed of people who have interest in news, politics, and automotive.

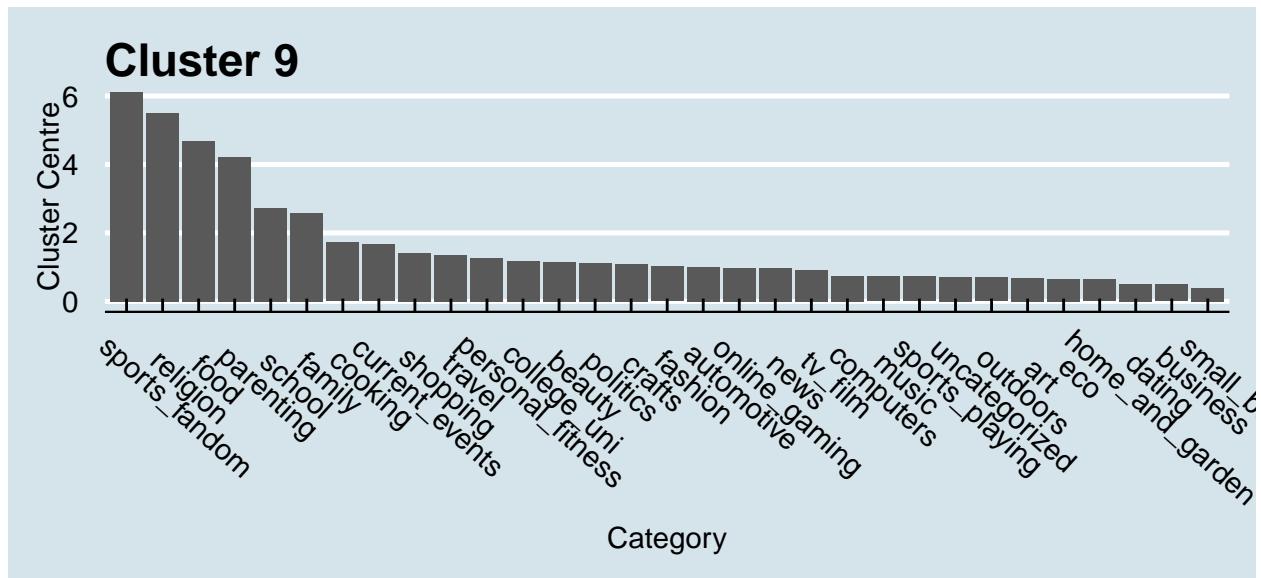


Figure 39: Cluster 9 primarily composed of people who have interest in politics, travel, computers.

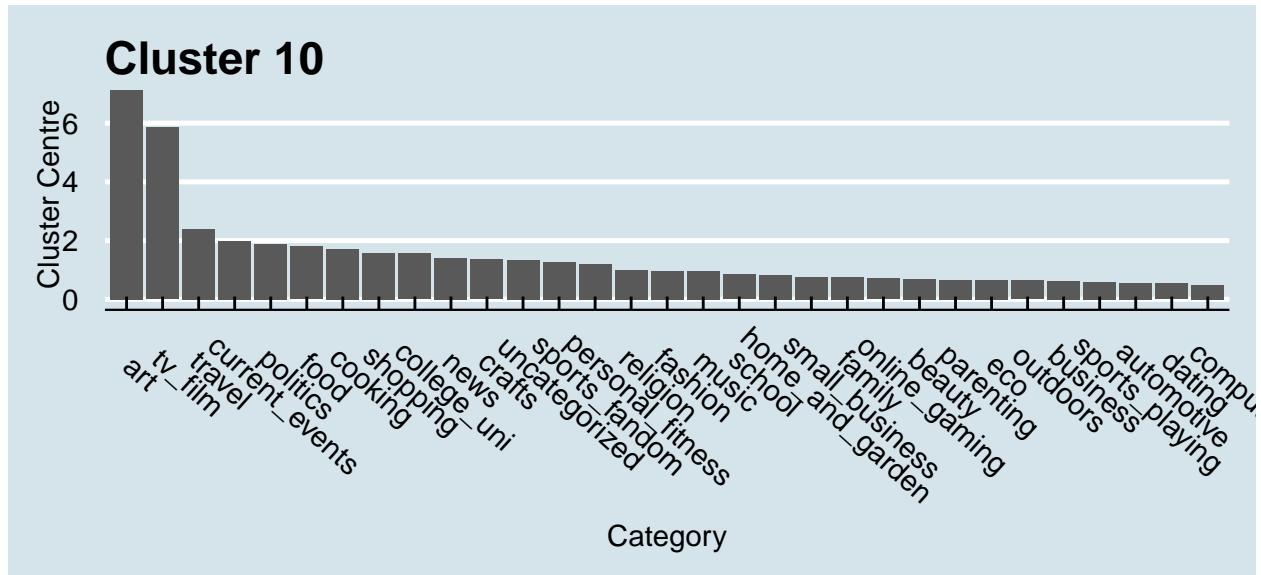


Figure 40: Cluster 10 primarily composed of people who have interest in personal fitness, cooking, outdoors.

K-Means clustering can allow us to identify the different market segments. In this case, we used 10 clusters to describe the unique market segments that the dataset has. These can then be used to create marketing campaigns.

## 8. Reuters Corpus

1. What question are we trying to answer - What topics does Bernard Hickey write about?
2. What statistical tool did we use to answer the question - We calculated tf-idf scores of each of the words in the document, then used those scores to perform a PCA and then clustered the documents based on similar topics and visualized it in a dendrogram.
3. Results: What evidence/results did your approach provide to answer the questions?
  - We obtained 5 major clusters and categorized all of Bernard's documents in those clusters.
  - Then we looked at the documents in each cluster and found what topics each cluster represents.
4. Conclusion: What are your conclusions about your questions? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set. - Bernard mostly writes about financials of companies, mergers and acquisitions, business and investing, politics and media.

### Preprocessing steps

1. Created a reader function to read documents.
2. Take out the second part of the file name and rename the files.
3. Create a corpus.
4. pre-processing/tokenization steps.
5. Removing stopwords.
6. Creating a doc-term-matrix.
7. Inspecting its entries.

```
## starting httpd help server ... done
```

## Basic summary statistics of documents

```
## <<DocumentTermMatrix (documents: 50, terms: 2547)>>
## Non-/sparse entries: 9419/117931
## Sparsity : 93%
## Maximal term length: 22
## Weighting : term frequency (tf)
```

## Inspecting the DTM

```
## <<DocumentTermMatrix (documents: 10, terms: 20)>>
## Non-/sparse entries: 61/139
## Sparsity : 70%
## Maximal term length: 22
## Weighting : term frequency (tf)
## Sample :
## Terms
## Docs adding aimed also annual asked august australia bank banks
##   1      1      2      4      3      1      1      1      3      3
##  10     0      0      2      2      1      0      1      0      0
##   2     0      0      5      0      0      1      0      0      0
##   3     0      0      6      2      0      1      1      0      0
##   4     0      0      4      1      1      1      1      0      0
##   5     0      0      1      1      1      0      0      0      0
##   6     0      0      1      1      1      0      0      0      0
##   7     1      0      2      1      0      0      1      2      0
##   8     0      0      5      1      0      0      0      1      0
##   9     0      0      5      1      0      0      0      1      0
## Terms
## Docs bernardhickeynewsmltxt
##   1          1
##  10         1
##   2          1
##   3          1
##   4          1
##   5          1
##   6          1
##   7          1
##   8          1
##   9          1
```

## Finding frequently used terms

Table 14: Frequent terms

x
also
bank
banks
bernardhickeynewsmltxt
character
datetimestamp
description
first

---

x  
growth  
heading  
hour  
isdst  
language  
listcontent  
market  
mday  
meta  
million  
min  
mon  
net  
origin  
percent  
profit  
profits  
said  
sydney  
wday  
will  
yday  
year  
australian  
corp  
ltd  
murdoch  
new  
news  
analyst  
analysts  
quarter  
half

---

**Further pre-processing steps** 1. Drop those terms that only occur in one or two documents. 2. Construct tf-idf weights.

## Comparing Documents

Here, we observe term weights in the first document.

```
## <<DocumentTermMatrix (documents: 1, terms: 935)>>
## Non-/sparse entries: 124/811
## Sparsity           : 87%
## Maximal term length: 22
## Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
## Sample             :
##   Terms
## Docs    aimed      cba    changed competitive     fall    falling
##       1 0.03499046 0.08747616 0.02863731 0.03499046 0.04711883 0.03499046
##   Terms
## Docs    industry    meeting  pressure    profits
```

```
##      1 0.05273955 0.02994768 0.0366789 0.03448276
```

Now, we compare this with the actual content of the first document.

Table 15: Document 1

---

x

The Commonwealth Bank of Australia (CBA) warned again on Tuesday of a subdued profit outlook for the current 1996/97 year and beyond, citing a competitive housing market, falling margins and higher wage costs. CBA chairman Tim Besley told shareholders at the bank's annual meeting that nothing in the first three months of the year had changed earlier forecasts of flat profits in the current year to June 30, 1997.

"We are now three months into the financial year and nothing has changed to prompt the bank to revise this assessment," Besley said in his chairman's address.

"1996/97 is shaping up to be every bit as challenging as we envisaged," he said.

"The home loan market remains intensely competitive."

Commonwealth Bank reported a net profit of A\$1.12 billion in 1995/96, up from A\$983 million in 1994/95, and said when it released its results in August that it expected earnings to be flat in 1996/97.

Besley said CBA's directors aimed to maintain the ratio of dividends per share to earnings per share at around 80 percent in 1996/97.

However, Besley also said later at the meeting that banking industry profits could fall over the long run and that CBA would have to work hard to ensure its profit falls were less than those of other banks.

"We have to be sure that ours (profits) come down by as little as any of the others," Besley said in answer to a shareholder's question.

Besley later told reporters banking industry profits would be under pressure in the future.

"The industry will be under pressure of falling profits unless, just like in any other industry, it pays great attention to its costs and gets very efficient," he said.

"There's no suggestion that (the CBA's) profits would fall off this year," he said, adding that it was the bank's objective to increase profits in 1997/98.

CBA Managing Director David Murray also played down Besley's earlier comments. "I read that comment to mean a fall in the growth rate in profits," he said.

When asked if he too aimed for a rebound from the flat 1996/97 to profit growth in 1997/98, Murray said: "Yes. It means marketing better and containing our costs."

But he said it was too early to be specific about a possible profit rise in 1997/98. Murray however also told the annual meeting that cost control would be difficult.

He said CBA had restricted wages growth to four percent in the two years before the most recently negotiated enterprise bargaining agreements.

"That (four percent over two years) is clearly not sustainable with the inflation rate and wage expectations where they are but we continue to bargain hard on that front."

CBA's share price closed down 11 cents at A\$11.90, having earlier fallen to A\$11.80 on the bearish comments from the annual meeting. "It's all on this talk of no earnings growth and wages pressure," one Melbourne broker said before the late session bounce.

Murray and Besley also repeated their view that official interest rates were more likely to be cut than raised before the end of the year.

- Sydney Newsroom 61-2 373-1800

---

## Dimensionality reduction

We performed PCA on the tf-idf scores.

```
##      Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.00000 0.05613 0.08599 0.10286 0.12905 0.73397
```



Figure 41: The first 2 principal components seem to retain most of the variance of the data.

### Loadings of PCA

We filter our dataset to represent each document with just 2 numbers from PCA.

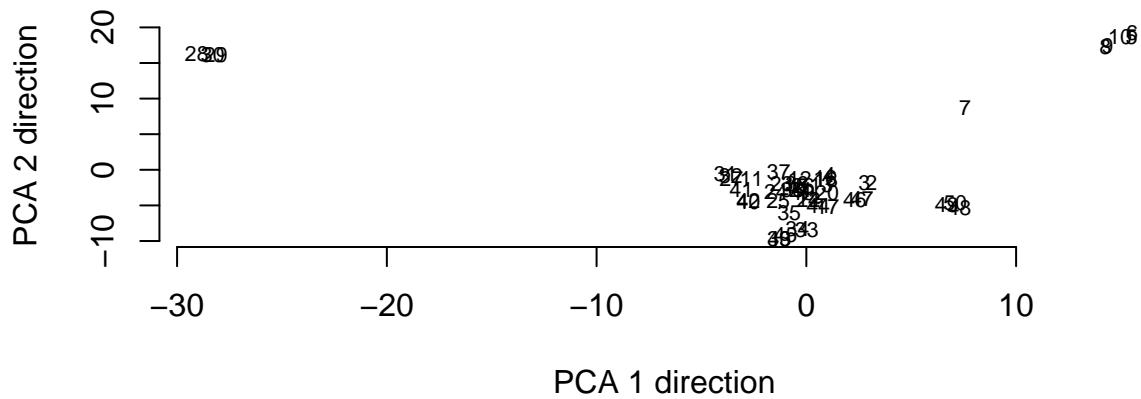


Figure 42: There are 5 clear clusters, most of the documents seem to fall in the third.

### Comparing documents in clusters

“Boosting revenue” topic - document 48 and 50

Table 16: Document 48

---

x

Rupert Murdoch's News Corp on Thursday reported a lower than expected profit for the first half of 1996/97, but later told analysts it remained confident of a 20 percent profit boost for the full year.

Strong earnings from the hit movie Independence Day and buoyant British newspaper sales helped drive net profits before abnormals 10.3 percent higher to A\$731 million (US\$555 million).

But the result was below analyst forecasts of a pre-abnormals profit of A\$735 million to A\$775 million.

"Everyone's a bit disappointed," said a Sydney analyst.

Andrew Sekely, the head of equities at broker Intersuisse, said: "It certainly wasn't a startlingly good result. The market has made its judgement". News' shares fell 16 cents to A\$6.60 after the result, but closed in Australia at A\$6.76.

"It's about A\$20 million less than what the market would have liked and to still reach that (20 percent profit growth) outlook, it's got to make up the difference in the second half," the Sydney analyst said.

News Corp had said in August last year when it posted a A\$1.26 billion net profit for the year to June 30, 1996 that it expected strong movie and television revenues to increase profits by at least 20 percent in 1996/97.

News Corp told analysts in a teleconference briefing on Thursday after the first half result that it still expected 20 percent profit growth for 1996/97.

"They did say that they're still on track for the 20 percent growth," another Sydney media analyst said after the briefing between company officials and U.S. and Australian analysts.

Analysts said News Corp officials were bullish about catching up in the second half and had pointed to a strong outlook for movies and U.S. television and newspapers.

Operating income from News' Fox Filmed Entertainment unit surged to A\$202 million for the first half of 1996/97 from A\$4 million a year earlier, due largely to *Independence Day*. || Analysts said there - release of *Star Wars* and other films would help further boost movie revenues in the second half. || British newspaper profits rose 1

---

Table 17: Document 50

---

x

Rupert Murdoch's global media group News Corp Ltd reported a lower than expected first half profit on Thursday, but later told analysts it remained confident of a 20 percent profit boost for the full year.

News said strong earnings from the hit alien movie Independence Day and buoyant U.K. newspaper sales helped drive up net profits before abnormals 10.3 percent to A\$731 million.

But this was below analysts' expectations of a pre-abnormals net profit for the half of A\$735 million to A\$775 million.

"Everyone's a bit disappointed," said one senior Sydney analyst who asked not to be named.

Andrew Sekely, the head of equities at broker Intersuisse, described the result as moderate.

"It certainly wasn't a startlingly good result. The market has made its judgment," he said. News' shares fell 16 cents to A\$6.60 in the hours after the result.

"It's about A\$20 million less than what the market would have liked and to still reach that (20 percent profit growth) outlook, it's got to make up the difference in the second half," the Sydney analyst said.

News had said in August last year when it posted a A\$1.26 billion net profit for the year to June 30, 1996 that it expected strong movie and television revenues to increase profits by at least 20 percent in 1996/97.

Murdoch repeated the forecast in October, but the group has been publicly tight-lipped about it since then and its first quarter results were also lower than expected.

However News reassured the analysts in a teleconference briefing after the first half result, saying that it was still on track for the 20 percent profit growth.

The profit forecast was not in the formal release.

"They did say that they're still on track for the 20 percent growth," another Sydney media analyst said after the morning teleconference briefing between company officials and U.S. and Australian analysts.

---

x

Analysts said News Corp officials were bullish about catching up in the second half and had pointed to the continued strong outlook for movies, U.S. television and newspapers.

Operating income from News' Fox Filmed Entertainment unit surged to A\$202 million for the first half from A\$84 million a year earlier, due largely to Independence Day.

"U.S. results were led by Fox Filmed Entertainment which posted a 154 percent gain in operating profits reflecting the continued success of Independence Day," News Corp said.

Analysts said the re-release of Star Wars, among others, would help further boost movie revenues in the second half.

"It's pure profit in the second half, whereas they had costs to write off in the first half," said another analyst. Strong revenues from News' Fox television network and its Fox stations would also help second half profits. "They did very well over the Super Bowl and they've got a new hit with King of the Hill," said an analyst. U.K. newspapers would again perform well in the second half, analysts said. British newspaper profits rose 18 percent in the first half, with gains in circulation revenues at The Sun, the Times and The Sunday Times. "All of the company's newspapers have maintained their dominant position in each of their respective markets, despite cover price increases instituted during the quarter at both The Sun and The Times," News said.

News said losses at its burgeoning Asian satellite broadcaster Star TV were in line with expectations but no figures were released. – Sydney Newsroom 61-2 373-1800

---

### **"Mergers/Acquisitions" topic - document 28 and 29**

Table 18: Document 28

---

x

An inquiry into Australia's financial system has left open the option of mergers among Australia's big banks and takeovers of local banks by foreign banks in an interim discussion paper released on Wednesday.

The inquiry, established by the government and headed by prominent businessman Stan Wallis, outlined various options for reform and was careful not to state its own preferences.

But it said the key aim of any reforms should be increasing the efficiency of Australia's banks to compete globally and that new technology would transform the sector. Big banks such as the National Australia Bank Ltd have used the same reasoning lobbying for relaxed merger rules.

The banks have said big bank mergers are necessary to compete globally and new technology such as the Internet and global competition meant such mergers would not cut competition.

Policy set by the former Labor government has stopped mergers or foreign takeovers of Australia's four largest banks and its two largest insurance and superannuation groups.

The current conservative coalition government has said this so-called "six pillars" policy will remain in place until it has considered the Wallis inquiry's recommendations, due to be delivered by the end of March 1997. "The Inquiry sees as its key goal the identification of means to increase the efficiency of the Australian financial system, without compromising its safety and performance," the paper said. Increased efficiency was needed to compete globally, to obtain the benefits of new technology and to increase investment returns, the paper said. It asked if competition and innovation could be stimulated by widening access to traditional banking activities through:

\* "Allowing direct non-bank access to the payments settlement system"

\* Better accommodating financial conglomerates

\* Allowing an increased range of institutions to provide a wider array of financial services, or

\* relaxing some of the ownership restrictions on financial institutions?"

The inquiry said it was also wanted any reforms to increase competition and asked in that context "whether there are any public policy grounds for restrictions on foreign acquisitions in the banking or insurance industries?"

---

x

In discussing the details of how bank mergers might be approved, the inquiry said there was widespread consensus that banks should be subject to the same competition rules as other industries. Currently any bank merger must be approved by the government through the Treasurer and by the Australian Competition and Consumer Commission (ACCC). This dual approval process does not apply in other industries.

"In undertaking its assessment, the Inquiry notes its preference... for all sectors of the economy to be subject to a uniform set of competition policy laws unless there is something special which justifies different treatment," the paper said.

The Inquiry said it would consider the following options for merger approvals, including;

\* keeping the current system of dual approval

\* having the Treasurer accept the ACCC's decision, or \* narrowing powers over mergers in banking and insurance law. It said there would be no need for a 'six pillars' policy if the Treasurer's powers were removed or if he agreed not to exercise them. "There was almost unanimous support that the 'six pillars' policy should be abolished," the paper said.

– Sydney Newsroom 61-2 9373-1812

---

Table 19: Document 29

---

x

An official inquiry into Australia's financial system said on Thursday it would look at recommending a relaxation of rules that currently stop big bank mergers and local bank takeovers by foreign banks. The inquiry, which was established by the conservative government after its March election, outlined various options for reform in a wide-ranging discussion paper and was careful not to state its own preferences. But it noted majority industry support for abolishing the rules stopping mergers and foreign bank takeovers and said it preferred an anti-monopolies system which treated the banking sector the same way as every other sector.

Until now the government has had the final say on bank mergers, rather than just the anti-monopolies watchdog.

Analysts said the tone of the discussion paper reinforced widespread expectations that the inquiry's final report would recommend allowing bank mergers and new foreign investment.

"It hasn't changed perceptions that ultimately the inquiry will make recommendations that allow the ACCC (Australian Competition and Consumer Commission) to permit bank mergers which haven't been permitted before," said ABN AMRO Hoare Govett Banking analyst Michael Pulman.

Shares in the big banks seen vulnerable to takeover have rallied since the beginning of the year in expectation that a conservative Australian government would allow the mergers.

Those seen as possible takeover targets include the Australia and New Zealand Banking Group Ltd and Westpac Banking Corp Ltd.

The National Australia Bank Ltd, Australia's largest and most profitable bank, is seen as the main predator. A policy set by the former Labor government and known as the "six pillars" policy has stopped mergers or foreign takeovers of Australia's four largest banks and its two largest insurance and superannuation groups. The current conservative Liberal-National coalition government has said this policy would remain in place until it has considered the Wallis inquiry's recommendations, due to be delivered by the end of March 1997. The inquiry, headed by prominent businessman Stan Wallis, said in its 415-page paper that any reforms should increase the efficiency of Australia's banks to compete globally.

It also said new technology would transform the sector.

Big banks such as the National Australia Bank have used the same reasoning when lobbying for relaxed merger rules.

These banks have said big bank mergers are necessary to compete globally and new technology such as the Internet and global competition meant such mergers would not cut competition.

"The inquiry sees as its key goal the identification of means to increase the efficiency of the Australian financial system, without compromising its safety and performance," the paper said.

---

x

Increased efficiency was needed to compete globally, to obtain the benefits of new technology and to increase investment returns, the paper said.

It said it would consider, “better accomodating financial conglomerates” and “relaxing some of the ownership restrictions on financial institutuions?”

The inquiry said it also wanted any reforms to increase competition and asked in that context “whether there are any public policy grounds for restrictions on foreign acquisitions in the banking or insurance industries?”

---

#### **“Australia-Murdoch media coverage” topic - documents 9 and 10**

Table 20: Document 9

---

x

Media magnate Rupert Murdoch said on Tuesday that the first quarter performance of his global media group, The News Corp Ltd, had been below expectations, but that he was still confident of a 20 percent profit rise in 1996/97.

“I am on record as saying that we expect a 20 percent increase in profit for the year,” Murdoch told a packed News Corp annual meeting in Adelaide.

“We still expect that and are still aiming for that during the coming year. However, I should say that the first quarter may not be quite up to those expectatations, but we will certainly be striving to make up any shortfall,” he said.

In the 1995/96 year ended June 30, News Corp’s net profit slipped to A\$1.02 billion (US\$806 million) from A\$1.37 billion in 1994/95.

He said revenues at the Fox U.S. television business had begun 1996/97 slowly because of of the Atlanta Olympic Games, the rights to which were held by a rival network.

Australian newspaper revenues would be flat in 1996/97, he said.

However, News Corp’s British newspapers were doing extremely well, with circulation at The Sunday Times climbing with little promotion while advertising in Britain was booming.

Murdoch, News Corp’s chief executive and chairman, also announced the group now had US\$2.5 billion in cash in the bank after recent U.S. bond issues and would leave it there.

He also said that News now planned to fund future expansion from cash flow where possible.

“We know that we have to be not only viable and profitable, but also that we should stay strong and liquid,” Murdoch said.

“As far as possible, we will finance them (News’ expansion plans) out of current cash flow to maintain liquidity so that we will be ready to take new opportunities as they arise,” he said.

News has often funded its aggressive growth in pay television and other media through either hefty debt or equity issues, which have sometimes met shareholder resistance.

Murdoch also announced that News planned to float its British based digital media technology company, Digital Media Services, within the next two to three weeks and would sell 20 percent of the company.

Digital Media Services, a combination of News Data Comm and Digi-Media Vision Ltd, would operate in Britain and Israel and produce technology for digital television.

He also later said News’ Asian satellite pay television operation, STAR TV, was performing well.

A\$1 = US\$0.79

---

Table 21: Document 10

---

x

Rupert Murdoch on Tuesday reported that News Corp Ltd had performed below expectations in the first quarter, but he said the global media group was still on track for a 20-percent profit rise this year.

Murdoch, the chairman and chief executive of News Corp, also unveiled a more conservative approach to financing new acquisitions.

"I am on record as saying that we expect a 20-percent increase in profit for the year," Murdoch told a packed News Corp annual meeting in Adelaide.

"We still expect that and are still aiming for that during the coming year. However, I should say that the first quarter may not be quite up to those expectations, but we will certainly be striving to make up any shortfall," Murdoch said.

Analysts viewed the comments as disappointing and said they were now more sceptical about the 20-percent profit pledge.

"If they can't meet expectations in the first quarter, when they have those strong revenues from Independence Day, then what hope do they have of getting 20 percent in the full year?" asked one Sydney-based media analyst referring to a hit Hollywood movie made by News Corp's Fox studio.

Traders on the Australian share market did not like the first quarter comment either, selling the shares 14 cents or almost two percent lower to close at A\$7.20, while the broader Australian equity market closed at a new record high.

In the 1995/96 year ended June 30, News Corp's net profit slipped to A\$1.02 billion (\$806 million) from A\$1.37 billion in 1994/95.

Murdoch said revenues at the Fox U.S. television business had begun 1996/97 slowly because of the Atlanta Olympic Games, the rights to which were held by a rival network.

Australian newspaper revenues would be flat in 1996/97, said Murdoch.

But he remained buoyant about News' British operations, saying the advertising market there was booming and the partly-owned BSkyB pay television operators was growing strongly.

Murdoch then unveiled a more cautious strategy for financing future expansion.

"As far as possible, we will finance them (News' expansion plans) out of current cash flow to maintain liquidity so that we will be ready to take new opportunities as they arise," he said.

News has often funded its aggressive growth in pay television and other media through either hefty debt or equity issues, which have sometimes met shareholder resistance. It almost collapsed under a debt mountain in the early 1990s.

Analysts said the comments indicated a more conservative approach towards acquisitions and their financing and would be welcomed by the market.

"It imposes a huge financial discipline on the company which would hearten people immensely," said another analyst.

Murdoch also announced that News planned to float its British-based digital media technology company, Digital Media Services, within the next two to three weeks and would sell 20 percent of the company.

Digital Media Services, a combination of News Data Comm and Digi-Media Vision Ltd, would operate in Britain and Israel and produce technology for digital television.

News Corp traditionally holds its annual meetings here in Adelaide, the South Australia city where Rupert Murdoch's father Keith founded the beginnings of the News Corp empire.

(A\$1 = \$0.79)

---

**Conclusion: even just these two-number summaries still preserve a lot of information**

## **Cluster documents**

Here we use hierarchical clustering to plot a dendrogram according to the PCA scores calculated earlier.

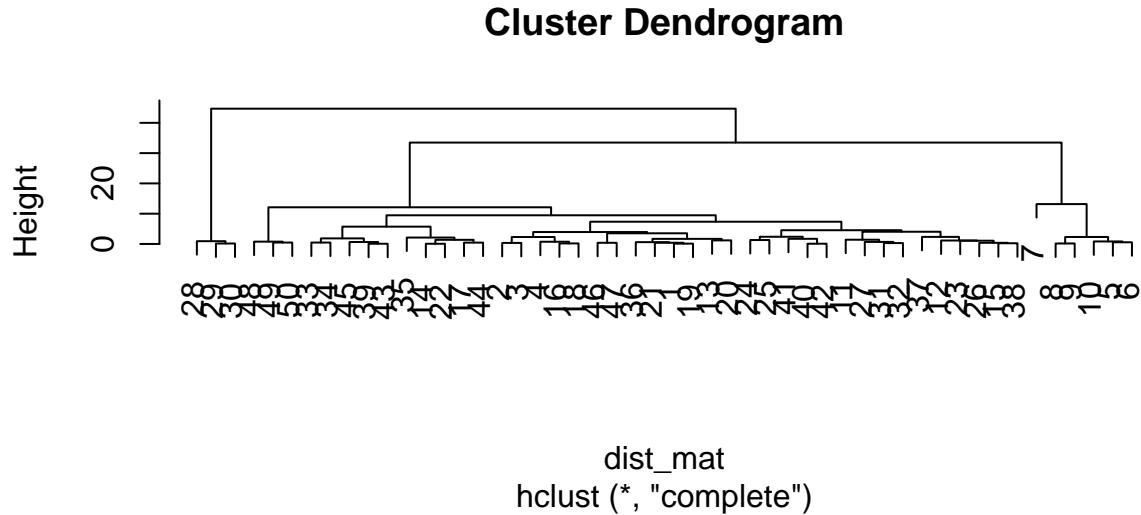


Figure 43: Document 7 doesn't seem to relate to any of the other documents.

## Inspect the clusters

Both documents belong to the Merger topic.

## 9. Association Rules Mining

### Approach

1. Before finding interesting association rules, we plotted the frequency of the top 10 items in all baskets. We can see that whole milk, other vegetables, and rolls/buns are the most frequent items in all baskets. For the apriori algorithm, we used a support of 0.005, confidence of 0.25 and then examined those that had a lift greater than 3. The support and confidence were chosen through trial and error of seeing the number of rules different values result in. A lift of 3 was chosen in order to only examine rules that lead to a higher probability of occurrence. After obtaining the rules using these thresholds, we visualized them in the network shown below. The areas of concentrated blue where whole milk appears to be the central node to a variety of vegetables and fruits is representative of the fact that these grocery items are frequently purchased together and with many other items (other nodes).
2. To look further into this, we sorted the rules by confidence and lift separately. We found that any time vegetables, fruit, and dairy products are purchased, the lift for whole milk was greater than 2.5. This means that the probability someone purchases whole milk increases if they are purchasing fresh produce and dairy products. We also see that the probability of purchasing vegetables increases if someone is already purchasing vegetables of a different kind. With respect to all dairy products, the probability of purchasing them increases when someone is purchasing fruit, which could be indicative of people making fruit smoothies or putting fruits in their yogurt.
3. Overall, the top 30 rules sorted by confidence and lift make sense since these groceries are considered essential to cook any meal and some are even complementary goods (i.e. fruit and yogurt, rolls/buns and sausage). Additionally, these rules can be used by grocery stores to increase sales by placing these products near each other. In fact, they are probably already using these rules to their advantage!

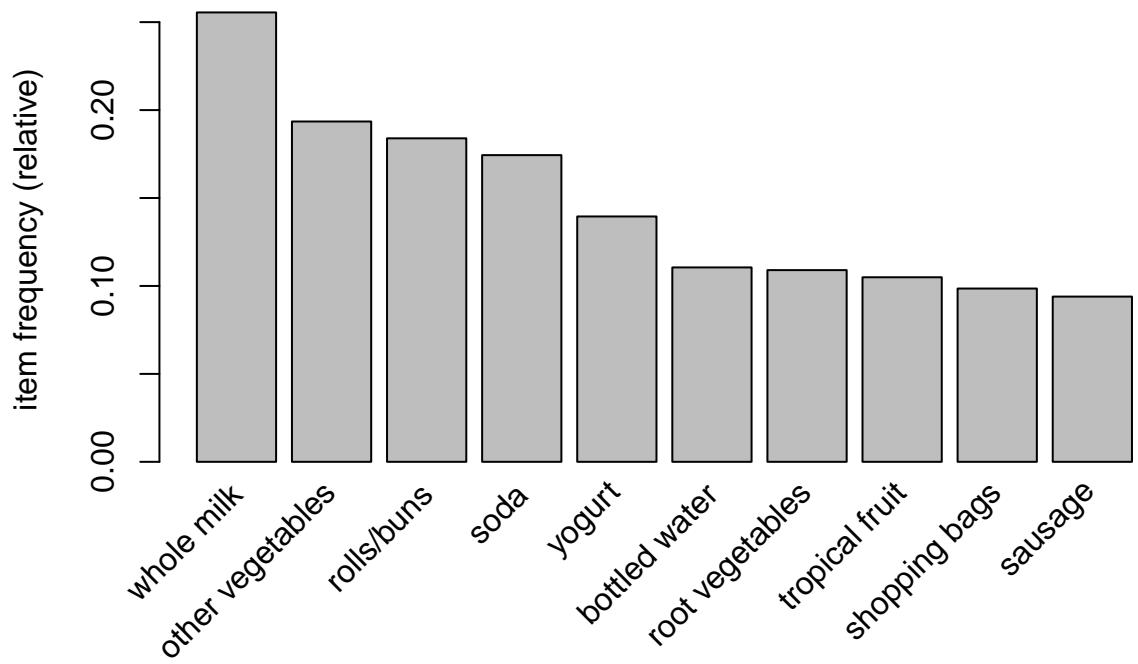


Figure 44: Plot of the item frequency of groceries in all baskets. Whole milk, other vegetables, and rolls/buns are the most frequently occurring items in baskets.

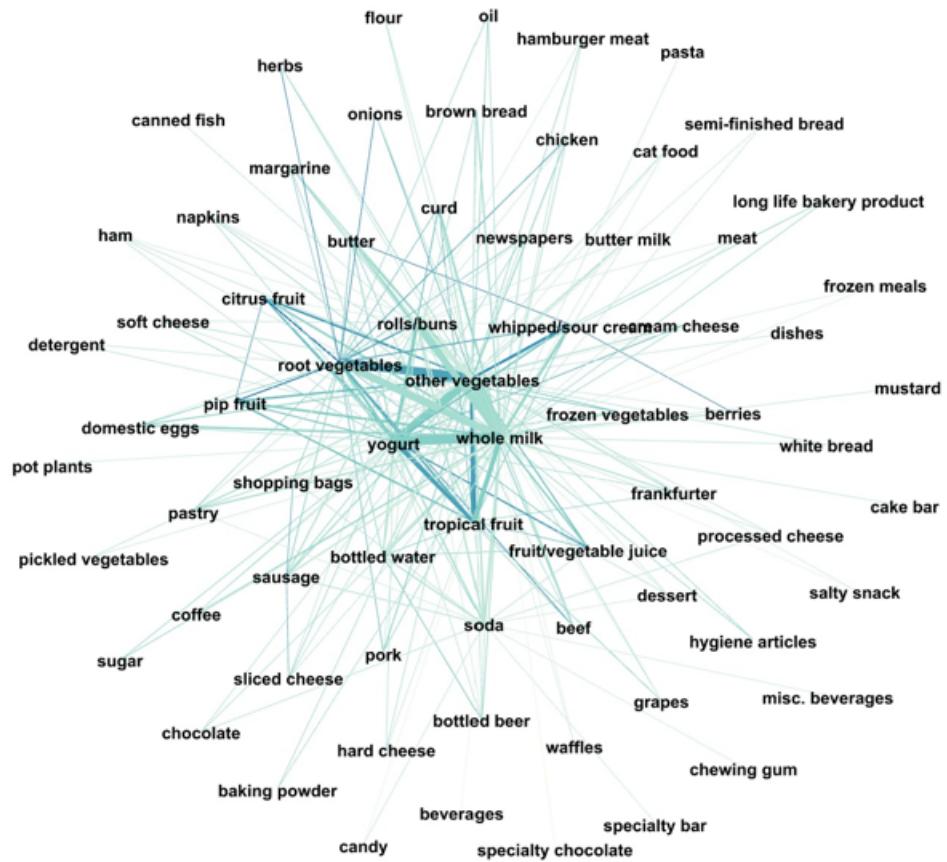


Figure 45: Network of grocery rules that have a support of 0.005, confidence of 0.25, and lift greater than 3. Concentration of blue coloring near whole milk and various fruit and vegetables indicates a co-occurrence pattern.

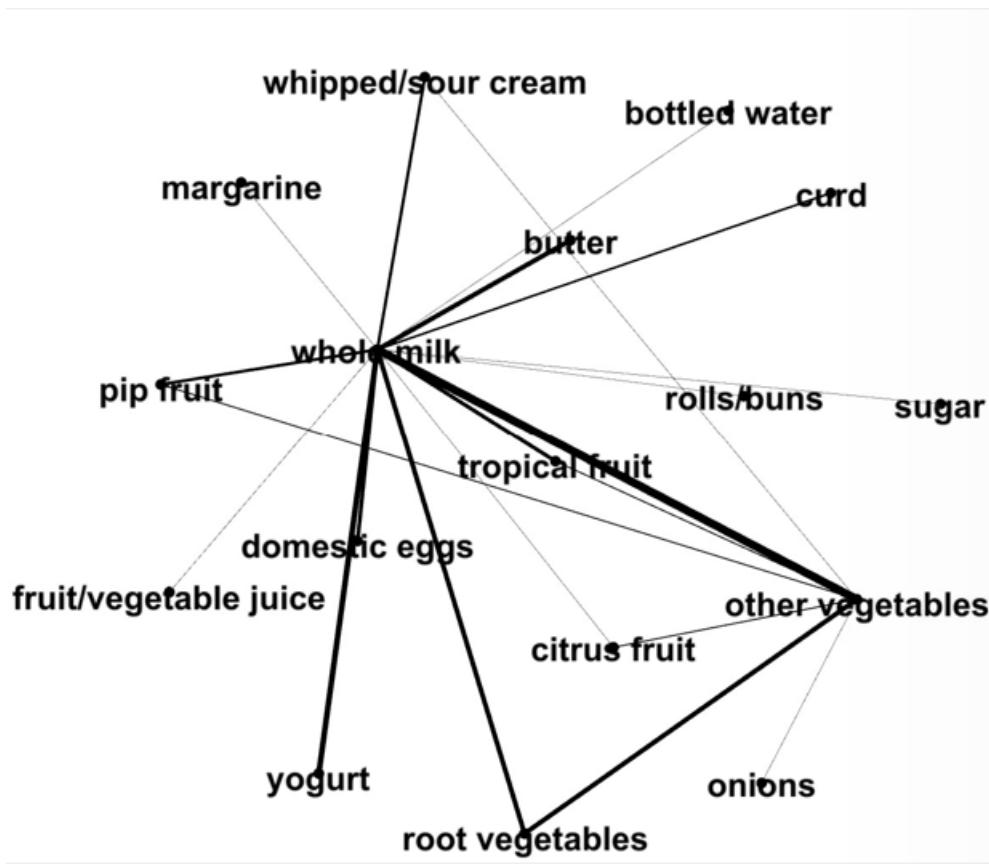


Figure 46: Network of the top 30 grocery rules sorted according to confidence. Displays the idea that the probability of purchasing whole milk increases with the purchase of fruit and vegetables.